Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Viet Ba Hirvola

Content Discovery in Online Services:
A Case Study on a Video on Demand System

Master's Thesis
Espoo, May 20, 2019

Supervisor: Professor Marko Nieminen, Aalto University
Advisors: Antti Salovaara, Ph.D
Einari Kanerva, M.B.A.

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

ABSTRACT OF MASTER'S THESIS

Author: Viet Ba Hirvola

Title:
Content Discovery in Online Services: A Case Study on a Video on Demand System

Date: May 20, 2019 **Pages:** 105

Major: Computer Science **Code:** SCI3042

Supervisor: Professor Marko Nieminen

Advisors: Antti Salovaara, Ph.D
Einari Kanerva, M.B.A.

Video-on-demand services have gained popularity in recent years for the large catalogue of content they offer and the ability to watch them at any desired time. Having many options to choose from may be overwhelming for the users and affect negatively the overall experience. The use of recommender systems has been proven to help users discover relevant content faster. However, content discovery is affected not only by the number of choices, but also by the way the content is displayed to the user. Moreover, the development of recommender systems has been commonly focused on increasing their prediction accuracy, rather than the usefulness and user experience.

This work takes on a user-centric approach to designing an efficient content discovery experience for its users. The main contribution of this research is a set of guidelines for designing the user interface and recommender system for the aforementioned purpose, formulated based on a user study and existing research. The guidelines were additionally translated into interface designs, which were then evaluated with users. The results showed that users were satisfied with the proposed design and the goal of providing a better content discovery experience was achieved. Moreover, the guidelines were found feasible by the company in which the research was conducted and thus have a high potential to work in a real product.

With this research, I aim to highlight the importance of improving the content discovery process both from the perspective of the user interface and a recommender system, and encourage researchers to consider the user experience in those aspects.

Keywords: video on demand, TV, recommender system, user experience design, user-centered design

Language: English

Acknowledgements

First of all, I would like to express my sincere gratitude to Liisa Puurunen, Mika Peuralahti and Einari Kanerva for giving me the opportunity to work on such an interesting project and for entrusting me with challenging tasks and responsibilities. Working at Elisa has been an invaluable experience, during which I grew as a professional and was able to apply the knowledge gained at the university in practice, within a real-world agile environment. Moreover, I was able to work closely with a team of great professionals: Lauri Rauhanen, Tuomas Tallqvist and Tero Soukko, who helped and supported me on many occasions. Furthermore, I would like to thank the Viihde team for insightful discussions, feedback and giving me the opportunity to continue my research.

My sincere thanks also go to Antti Salovaara and Marko Nieminen for their guidance, extremely helpful feedback, immense knowledge and positive attitude. I could not have imagined better mentors for my thesis. I would like to also thank Janin Koch for sharing her tips and experience in research and studying design.

Last but not least, I would like to dedicate big thanks to my family for supporting my decision to pursue my career in Finland, my wonderful friends Asta Korkeamäki, Salla-Marja Hätinen, Janin Koch, Martyna Szymczak, Yunfei Xue, Yichi Liao, and most importantly – my amazing husband Mikko Hirvola. They have given me the strength to work hard and reach my goals, before and throughout the thesis work, by listening to my troubles whenever I needed it, always believing in me and genuinely cheering for me.

Espoo, May 20, 2019

Viet Ba Hirvola

Abbreviations and Acronyms

RQ	Research Question
VoD	Video on Demand
PoC	Proof of Concept
Viihde	Elisa Viihde
UX	User Experience
UI	User Interface
HCI	Human-Computer Interaction
CF	Collaborative Filtering
CBF	Content-Based Filtering

Contents

List of Figures

Chapter 1

Introduction

Watching TV is an important activity during leisure time in human lives. For instance, between 1949 and 2010 its consumption almost doubled in the USA: from 4 hours 35 minutes per day to 8 hours 55 minutes [1]. The average daily TV watching time has been decreasing in the 2010s, outrun by the growing popularity of the Internet usage. Nonetheless, it still occupies most of the daily media consumption worldwide, and almost as much as the Internet usage in Finland (see Figure 1.1[2]).

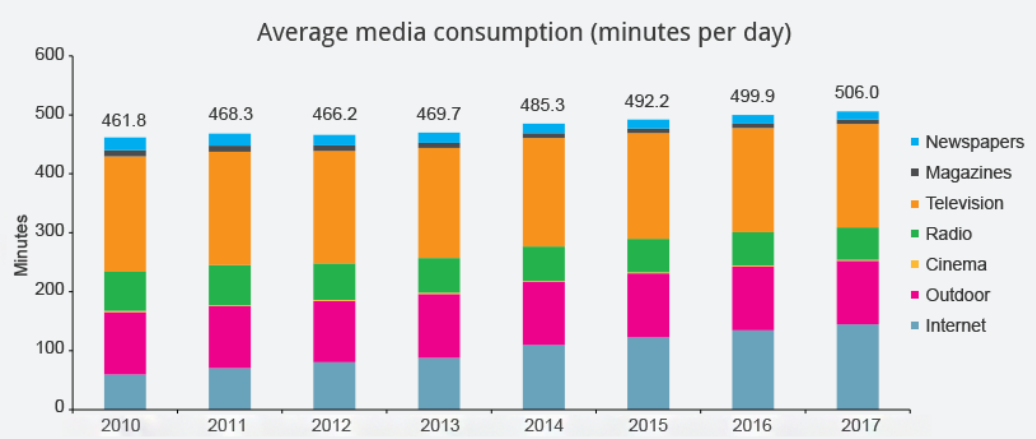
TV watching became a popular leisure activity, because TV is a universal device which does not require much effort (e.g. in comparison to going outdoors or dressing up for events) nor abilities [3, 14, 21]. The major motivations to watch TV are hedonistic, e.g. for entertainment and self-escape reasons, but also more related to seeking insight, such as self-development and self-reflection [86].

The growing availability of the Internet, increasing demands for consuming video content and enlarging consumer population resulted in the popularity of online video-on-demand (VoD) services [88], such as Netflix, Amazon Prime or Hulu. This in return changed notably the viewing habits: instead of having to wait for the TV programme to be broadcasted, one can browse, select and watch desired content at any time. However, having more control and flexibility over what to watch imposes an important issue: while in the past decision making in terms of TV watching was limited mainly to the choice of the channel at a given time, nowadays users are faced with significantly more options in many different services, e.g. Netflix alone has over 5500 TV shows and movies in the USA and over 2000 in Finland[3].

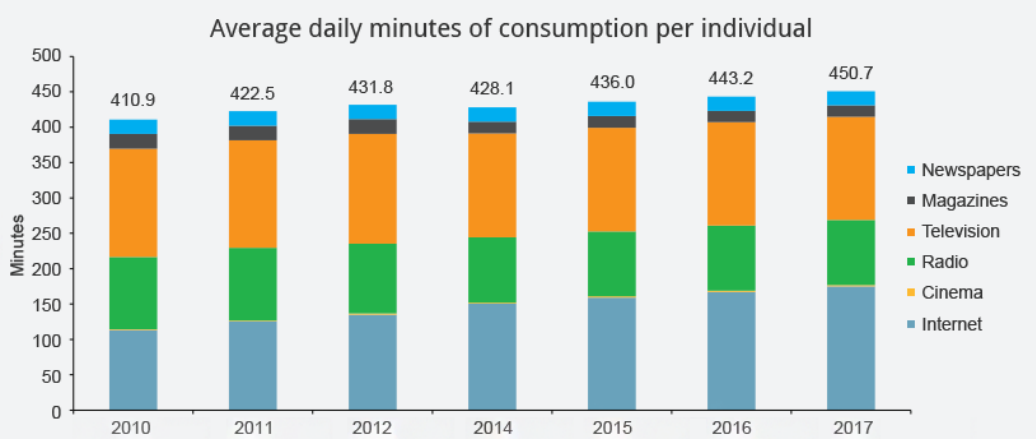
[1] When Did TV Watching Peak? https://www.theatlantic.com/technology/archive/2018/05/when-did-tv-watching-peak/561464/

[2] Media Consumption Forecast 2015. https://www.zenithmedia.com/

[3] Total of Netflix TV shows and movies by country. https://www.finder.com/



(a) Worldwide



(b) Finland

Figure 1.1: Media consumption worldwide and in Finland. (Source: Zenithmedia, font sizes enlarged by the author)

This thesis was conducted in Elisa Oyj, a Finnish company specialising in telecommunication and digital services [4]. The goal of the project was to create a proof of concept (PoC) of a video-on-demand (VoD) service. One of the main aims was to allow video content to be discovered efficiently. This research took on a user-centric viewpoint to ensure that the service can be used by a large population (*universal usability*) and be enjoyable to use as it is an entertaining service (*hedonic usability*) [31].

global-netflix-library-totals

[4] https://corporate.elisa.com/

1.1 Problem Statement

In the consumption psychology research, the situation in which users are overwhelmed with the number of available options is referred to as an *overchoice* or *choice overload*, which can lead to two phenomena having negative effects on both users and customers: *paralysis analysis* and *buyer's remorse* [56, 72]. The first phenomenon describes the situation in which the user over-analyses the problem and ends up not making a decision, and the latter refers to the feeling of regret after making a choice hastily [28, 78]. The overchoice problem is common in online services which require the user to choose among many options, e.g. in VoD or e-commerce systems, and has been discussed as the failure from a service to filter data meaningless for the consumer [56]. For instance, Netflix's typical user loses interest after 60 to 90 seconds of browsing, equivalent to going through 10 to 20 titles while stopping at around 3 to read in more detail [25]. Moreover, reduced browsing was found correlated to the perceived higher effectiveness of the system [40].

Research on improving content discovery in online services has been concentrated on building recommender systems, which aim to automatically suggest to the user a subset of content considered to be relevant based on their history of consumption and possibly other factors, e.g. [12, 41]. The main problem is that most of the research has been focused on improving the prediction accuracy of the algorithm, but less consideration has been made to its actual usefulness for the user and user satisfaction. Moreover, solving the problem of overchoice in terms of VoD services has not been extensively discussed from the perspective of the user interface and the interaction needed to browse the content. Therefore, this work focuses on the *user experience* aspect of content discovery. The overall research question (RQ) is thus:

How to design an efficient content discovery experience in the context of a VoD service?

Task completion time, often used in HCI research to evaluate efficiency, is not a suitable measurement for leisure activities, during which people are supposed to be relaxed and not rush their actions. Here, efficiency refers to the user's *perceived effort* spent on browsing content. In other words, the goal is to create an experience where the user can find content in a manner that keeps them satisfied with the service and with low *interaction cost*[5]. The output of this thesis are a set of design guidelines and concepts for creating such experience from a user-centered perspective, in the context of a VoD service.

[5]Interaction Cost. https://www.nngroup.com/articles/interaction-cost-definition/

1.2 Structure of the Thesis

The scope of the thesis is as follows. Chapter 2 firstly describes the project in which this research was carried out, then existing guidelines for designing interfaces for TV, common approaches to building recommender systems from the literature, and two state-of-the-art VoD systems (Netflix and Elisa Viihde). The research question is then reformulated given the background information. Chapter 3 describes a user study conducted to gather design implications. In the experiment, participants were observed while using Netflix and Viihde, after which they designed their ideal VoD service. The design implications are then transformed into UI design decisions and wireframes for the Home screen and a new browsing method referred to as a Category carousel. The design decisions for the proposed wireframes are discussed in Chapter 4. In Chapter 5, possible solutions satisfying the implications in terms of recommender systems are further explored from the literature, and the feasibility of data that could potentially be used for it is investigated. Chapter 5 is finalised with suggestions of potential solutions for implementing and designing the recommender system, based on findings made so far. Proposed designs and guidelines are evaluated by users and experts in Chapter 6. Finally, research findings summarised into 12 design guidelines, the limitations and future work are discussed in Chapter 7 and concluded in Chapter 8. The general workflow of this research is visualised in Figure 1.2 and annotated with the main chapter numbers in which its parts are discussed.

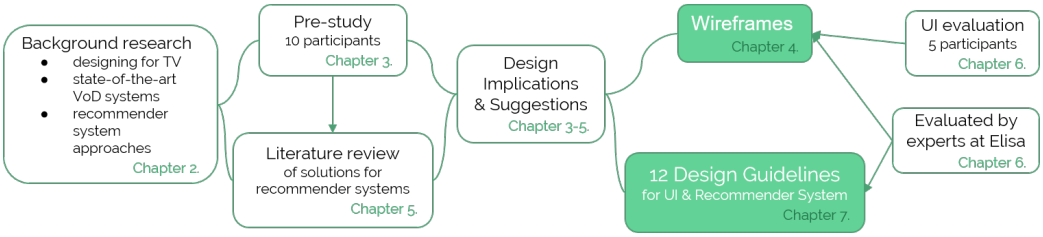


Figure 1.2: General workflow of the research.

Chapter 2

Background

This chapter firstly gives an overview of the project within which this thesis was carried out in Section 2.1. Next, design implications for TV are summarised in Section 2.2 and different approaches to building a recommender system are discussed in Section 2.3. Two existing VoD systems: Netflix and Elisa Viihde are described in Section 2.4. Given the context of the project and all background information, the chapter ends with reformulation of the original research question.

2.1 Project Environment

This research was carried out in collaboration with Elisa Oyj. The primary goal of the project was to create a proof of concept (PoC) VoD service for Android TV, which could be provided as a configurable framework for different operators to offer their video content in. The back-end (server side) was provided externally, thus the focus of this project was on developing the front-end (client side). The project started in November 2018 with a scrum team consisting of: the product owner (PO), 8 software developers and a visual designer. I joined the project in early January 2019 with the responsibility to conduct user research and re-design the UX and UI. One of the main requirements for the new interface was to allow the user to quickly discover content, which in turn became the RQ of this research. The PoC ended in the first week of April 2019, making time the main constraining factor in this work. Most user research activities affecting the front-end of the application had to be conducted within a 2-month timeframe to give the developers 1 month for implementing the interface. Remaining research activities were conducted after that, with the aim to gain more insight into the researched problem for future work.

2.2 Designing for TV

Unlike many other interactive devices, such as mobile phones or laptops, the usage of TV is substantially different due to the distance between the user and the device, typically spanning within multiple meters, depending on the size of the screen. This also implies that, in order to avoid having to frequently walk-up to the device, the user must mainly interact with it through a remote control. Traditional remote controls are equipped with a directional pad (D-pad), which means that on-screen navigation is possible through moving from one item to another using four directional buttons (up, down, left, right) [26]. This significantly limits user's movement within the system in contrast to other commonly used devices: the user can tap at any point on the screen in touch devices, move a cursor in all directions at different speeds in WIMP (Windows, Icons, Menus, Pointers) interfaces [39], but only navigate in two-axis with a D-pad.

Google provides its own design principles and guidelines for Android TV interfaces [26]. Based on those, the main considerations in designing for TV are the following:

- **Amount of text**. Reading from a distance may be difficult for users, therefore the amount of text should be limited and chunked for easier reading, if possible [15, 26].

- **Layout**. Since the user must be able to easily move between elements with a D-pad, the most common and natural layouts for TV interfaces are: a list for one-dimensional data; and a grid layout for browsing image-based content [13].

- **Alignment of objects**. Interactive objects should be aligned such that there is a clear two-axis path between them using, e.g. overlapping objects or nested hierarchies should be avoided (see Figure 2.1) [6, 15, 26].

- **Indicator of the current position**. Touch devices do not need a cursor since the user can simply touch the target area. In case of WIMP and TV interfaces, the input is provided through an external device (mouse and remote control, respectively), therefore a pointer to indicate the current position on the screen is needed [15]. On TVs, the currently active object acts as such indicator (later referred to as a *"focus point"*). Therefore the following must be met:

  – There must always be an object in focus [26].

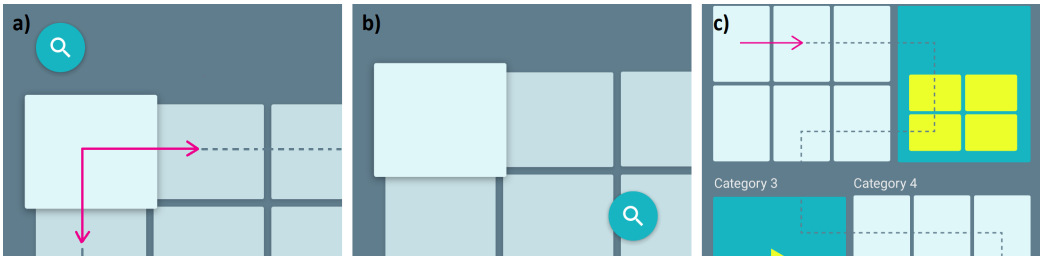


Figure 2.1: a) Alignment of objects allows navigating between them in a straight path. (b) The Search button is placed in between and on top of objects, making the path unclear. (c) The path between nested objects is complex. (Source: Google's AndroidTV Design Guidelines [26])

– The focus point must be clearly indicated by making it visually different than surrounding objects [26]. This is commonly done by making the object larger, adding shadows or drawing a border around it.

Moreover, TVs is often a shared device, making it necessary that systems for TV support multi-user usage, but also adapt to the preferences of an individual and respect the privacy of the user if they are not comfortable sharing what they have watched, as TV watching can be a "guilty pleasure" [3, 6]. Lastly, it is important to remember that traditionally watching TV is a leisure activity, hence TV applications should not feel like tools designed for complex tasks, but rather give a relaxed and enjoyable experience [3, 14]. This also implies that when evaluating a TV interface, traditional performance metrics and heuristics used in usability evaluations, such as task completion time or the number of clicks, do not reflect the laid-back nature of TV usage [14].

2.3 Recommender Systems

People receive recommendations everyday from different sources like friends, family, mentors or news. Thus, it is no surprise that in the digitalised world we live in there is a need for recommendations as well. Recommender systems were introduced to address the problem of overchoice and personalise their offers by suggesting a smaller, but more suitable for a particular user collection of items [28]. The recommendations are based on different data collected about the user as they use the service. Producing high quality

recommendations is beneficial for both users and companies who use them, since personalised content attracts customers.

Hence, nowadays it is expected that online services provide the user with personalised suggestions, whether they are products on Amazon [44], jobs on LinkedIn [37], friends in Facebook[1], videos on YouTube [17] or Netflix [25]. This section describes different approaches to building a recommender system and methods used in two state-of-the-art VoD services.

2.3.1 Rule-based Filtering

Early recommender systems used simple heuristics which define rules that determine what item is recommended to the user based on their past interaction with other items [61]. For instance, in the case of VoD systems, a rule could be to recommend other content from the same director or sequels of movies that were watched. Rule-based recommender systems are simple to implement using if-else statements and are good at capturing common heuristics people use when looking for items. The rules can be defined manually by experts, in which case the technique is called *knowledge-based filtering* (KBF) [28]. A more robust approach is to use a data mining technique called *association mining* or *association rule mining*, commonly used in e-commerce. It attempts to find associations between items that were purchased together [28, 70, 71].

However, rule-based recommender systems are poor at capturing individual variability between users. Moreover, since they work on a pre-defined set of heuristics that apply to all users, they do not offer high detail of personalisation [61].

2.3.2 Content-based Filtering

Recommender systems relying on the features ("content") of items or users are referred to as *content-based filtering* (CBF). Their training data typically consists of a list of items that the user consumed (e.g. watched in the context of VoD systems), and each item is defined by a set of features, e.g. a film genre. Therefore an item I can be represented with a vector of n features $I = (f_1, f_2, ..., f_n)$. [83] An example of an early implementation of CBF is a recommender system described by Foltz and Dumais (1992) [20], which used semantic information extracted from technical memos as features to suggest relevant memos to the employees at Bellcore.

[1]Where do People You May Know suggestions come from? https://www.facebook.com/help/163810437015615

CBF methods are usually implemented with classification algorithms, which are trained on data divided into categories whether the user liked an item or not, and then classifying a new item either as positive (recommend) or negative (do not recommend) [61, 83]. Identification, whether the user likes an item or not, can be done through so called *explicit feedback*, such as user rating, or *implicit feedback*, by inferring that through analysing their past behaviour and interactions with the items [61, 74]. CBF is suitable for the *cold start* problem, which defines the inability of a system to address new items or users due to insufficient data gathered about them, because it relies on their features which are mostly known [34]. However, because of that, CBF systems often tend to recommend very similar items [53, 74], and are unable to identify underlying and complex relations between items and users that are not explicitly related to their features [10]. Moreover, to improve the quality of recommendations, it is often needed to collect external information, which may be unavailable or difficult to obtain [34].

The main challenge in CBF is often caused by the type of data that describes the features. The algorithm needs to be able to analyse it, which is often difficult, e.g. for *unstructured data* such as free text [10]. Instead, *structured data* such as numerical, binary and nominal attributes are the easiest for a machine to analyse and train on, therefore many techniques to transform unstructured data into structured representations were researched.

A common way to convert unstructured text to structured data is to create a binary indicator that a term of interest is present in the text [61], e.g. if we are interested whether a movie has violent scenes, we could create a value representing whether terms related to violence (e.g. assassination) are present in the plot description. One of the commonly used and simplest techniques is *tf-idf*, which stands for term frequency-inverse document frequency, which aims to describe the importance of a term in a textual document based on how often it occurs in a particular text and the entire collection [53]. Many other techniques to analyse unstructured data has been introduced in the fields of text mining and natural language processing, however, it is still a great challenge due to the complexity of a natural language (e.g. one word can have different meanings depending on the context), the nature of use (e.g. sarcasm) and diversity of languages.

Therefore, the choice of the algorithm highly depends on the type of data that features are represented with. For instance, decision trees and Euclidean distance are suitable for structured data, cosine similarity for vector spaces or Naïve Bayes for text-based data [61].

2.3.3 Collaborative Filtering

Current recommender systems are most often based on collaborative filtering (CF) [71], e.g. it is used by a popular e-commerce service Amazon [77]. Unlike CBF, CF approaches do not rely on the features, but on the past behaviour of a user and users-item ratings, to deduce unknown relationships between them [34, 74, 77]. GroupLens project by Konstan et al. (1997) [41] was one of the first to utilise this approach. Since 2006, CF has gained notably more interest after Netflix launched "The Netflix Prize"[2] competition for the best algorithm to make predictions based on a user's past behaviour and their similarity with other users [42]. Generally, CF has been found more accurate than CBF algorithms, but more vulnerable to the cold start problem [34]. CF methods can be divided into *memory-based* and *model-based* algorithms.

Memory-based CF is easy to implement and generally attempts to calculate similarities either between users (user-based) or items (item-based). The main idea of user-based filtering is that users with similar behaviour within the service (e.g. watched similar videos) will like the same items and are thus recommended items they have not consumed yet, which other like-minded users did. In the item-based approach, the idea is that items with a similar consumption history will be liked by users who already consumed one of them. [10, 73, 85] Item-based approach was found to improve the quality of recommendations and was more preferred by users in many cases [34]. Methods commonly used for CF are similarity measures (e.g. cosine similarity), correlations (e.g. Pearson's, Spearman's) or simple measures like a weighted average [77]. CF systems can be used to find top-N recommendations, for instance using k-nearest neighbour (KNN) algorithm to find k most similar users and then aggregate N most frequently consumed items within them [45, 77].

The data used in CF is usually a list of n users $u_1, u_2, ..., u_n$ and m items $i_1, i_2, ..., i_m$, and each user has a list of items to which their feedback is known or inferred, that is represented in a form of a user-item matrix [77]. A simple example of such matrix is shown in Table 2.1.

The main disadvantage of memory-based CF approach is that while they perform well on smaller and dense datasets, the results become unreliable when there is a lot of missing data. Moreover, they do not scale well, because they need to process the whole user-item matrix to compute similarities between all items and users, which is computationally expensive. [10, 74, 77] This degrades the usefulness of CBF in real systems. To solve the scalability

[2]The Netflix Prize. https://www.netflixprize.com/

	i_1	i_2	i_3	...	i_m
u_1	+			...	-
u_2		+	+	...	-
u_3	+	+			
...	...	...		...	...
u_n	-	+		...	-

Table 2.1: User-item matrix of n users and m items. "+" indicates that the user liked the item, "−" that they did not. Otherwise, the data about the user-item relationship is empty (missing).

problem, multiple techniques have been introduced, such as dimensionality reduction with Singular Value Decomposition (SDA), Principal Component Analysis (PCA), or *matrix factorization*, however, as a result, the quality of predictions may be reduced [73].

To address the limitations of memory-based CF algorithms, machine learning and data mining models have been investigated. In model-based CF algorithms, the model is constructed to represent the user's rating behaviour and obtain underlying characteristics to predict ratings of items that were not interacted with [10]. Probabilistic methods such as Bayesian models have been found to deal better with sparse data. Clustering methods allow finding recommendations within the clusters instead of the whole database. However, model-based CF is generally much more complex and expensive to build (e.g. due to a high number of parameters that need to be tuned), and sensitive to changes in data. [10]

Lastly, CF algorithms are vulnerable to so called *shilling attacks*, which are spam attacks from users who intend to misdirect system's recommendations by giving a lot of positive feedback on items they are affiliated with [10, 77].

2.3.4 Hybrid Approaches

To reduce the limitations of CBF or CF, multiple hybrid approaches have been proposed. Melville et al. (2002) [47] introduced a *content-boosted collaborative filtering* approach which uses bag-of-words Naïve Bayes CBF algorithm to fill in missing data by predicting unrated items and then provide recommendations with pure CF based on the less dense data.

Another approach to combine the methods was presented by Mortensen et al. (2008) [53], who created a hybrid recommender system where users initially are recommended items using CBF to address the cold start problem, and after a week a CF user-based system using mean-squared difference

measure is utilised instead.

2.4 State-Of-The-Art VoD Systems

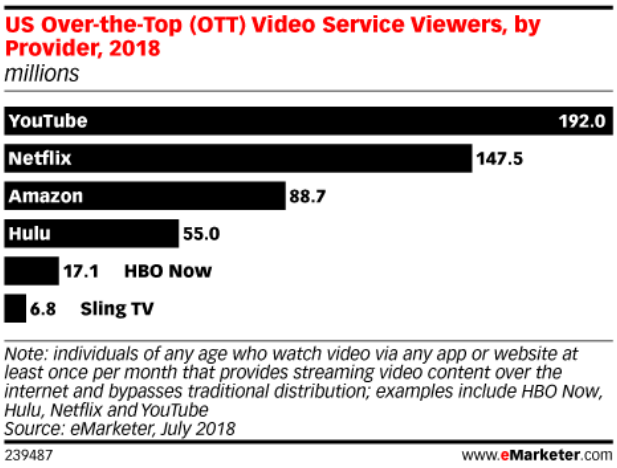


Figure 2.2: Comparison of the amount of viewers between over-the-top (OTT) video service providers. (Source: eMarketer)

As shown in Figure 2.2[3], among the largest video streaming services based on the number of users are YouTube and Netflix, with 192 million and 147 million users in 2018, respectively. YouTube, unlike the others on the list, allows users to upload the content to the service, therefore the amount of content is significantly higher and the metadata is very poor. Moreover, most of the videos on YouTube are short and of varying quality, therefore making the user interactions and intentions different than in the case of Netflix. [17]

2.4.1 Netflix

Netflix is one of the most widely used VoD systems and is available in 190 countries [4] as of Q1 2019. It typically displays on one page about 40 rows of videos (later also referred to as carousels) with 75 videos per row, which may vary depending on the capacity of the used device [25]. This means that a

[3]US Over-the-Top (OTT) Video Service Viewers, by Provider, 2018 (millions). https://www.emarketer.com/Chart/US-Over-the-Top-OTT-Video-Service-Viewers-by-Provider-2018-millions/220561

[4]Where is Netflix available? https://help.netflix.com/en/node/14164

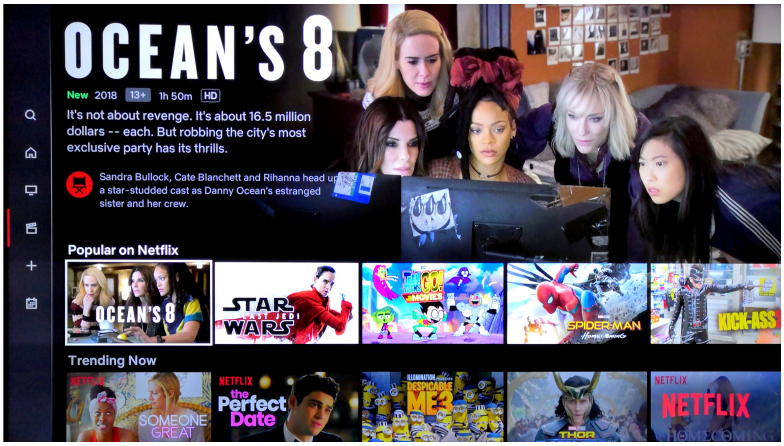


Figure 2.3: Film screen in Netflix (Android TV version).

user has 3000 (possibly duplicated) choices on one page. Figure 2.3 shows a screenshot from the Film page on Netflix. At the top there is a block with a description of the currently selected video and a large thumbnail on its right side, while at the bottom of the screen there are two carousels with videos displayed at a time. The description block typically includes metadata about the year of release, age restriction, duration, video resolution, plot description and cast. Moreover, for videos which have been long enough in the service, there is a percentage number indicating how much the system estimates a particular video to match user's interests. In the video selected in Figure 2.3, there is still a "New" indicator instead of the percentage.

Gomez et al. (2016) [25] from Netflix describe the recommendation system used by the service, mentioning that different mathematical, statistical and machine learning algorithms are used, both supervised and unsupervised learning approaches, which consider different time windows and popularity. While it is not clear what methods are exactly used, we can learn from their paper that Netflix utilises different algorithms for different purposes:

1. Carousels based on genres are personalised to the user, therefore each user may see different videos in the same genre.

2. Top-N carousels display best picks personalised for the user from the entire video catalogue.

3. Trending carousels display the most popular videos based on short-term trends.

4. Because-you-watched carousels contain results of a video-video similarity calculated for every video and is not personalised.

5. In the continue watching carousel, an algorithm attempts to infer whether a user intends to resume the video or was it abandoned.

Based on the descriptions above, it is likely that algorithms 1, 2 and 4 are CF algorithms, while 3 and 5 could be rule-based. Additionally, which rows are shown to the users is also personalised with the focus to offer diversity needed to target different use cases and contexts. The authors reported that until 2015 the row selection was done with a rule-based algorithm. To initially train the recommender system, Netflix shows during profile creation a grid of thumbnails with different films and TV programmes and asks a new user to pick 3 videos from it.

Furthermore, 80% of the content is discovered from browsing and 20% from search. For every search query, the system tries to provide alternative results. Netflix also changes regularly what content, carousels and thumbnails of videos are displayed to the user.

Lastly, Gomez et al. (2016) [25] also discuss the challenge of testing a recommender system. The authors describe that it is a time consuming process, which typically takes between 2 to 6 months.

2.4.2 Elisa Viihde

Elisa Viihde[5] is a video streaming service which includes multiple sub-services within it, such as: live and on-demand content, in-built connection to external television services like HBO Nordic[6] or C More[7]. According to the most recent Elisa Annual Report[8], Viihde had over 400 thousands customers in Finland and Estonia. In a survey by Statista in 2018 on video streaming services in Finland[9], Netflix was the most popular service with 37% respondents currently subscribed to it, followed by Elisa Viihde's 11%.

Later in Chapter 3, an Android TV version of Viihde will be investigated. It is currently a beta version, therefore it does not contain all features and functionalities oh Elisa Viihde. In the Android TV version, the on-

[5] https://elisaviihde.fi/

[6] https://fi.hbonordic.com/

[7] www.cmore.fi

[8] Elisa Annual Report, 2018. https://corporate.elisa.com/attachment/elisa-oyj/annual-report-2018/Elisa_vk18_annual_review.pdf

[9] Online video service subscriptions in Finland 2018, by platform. https://www.statista.com/statistics/744246/survey-on-online-video-service-subscriptions-in-finland-by-platform/

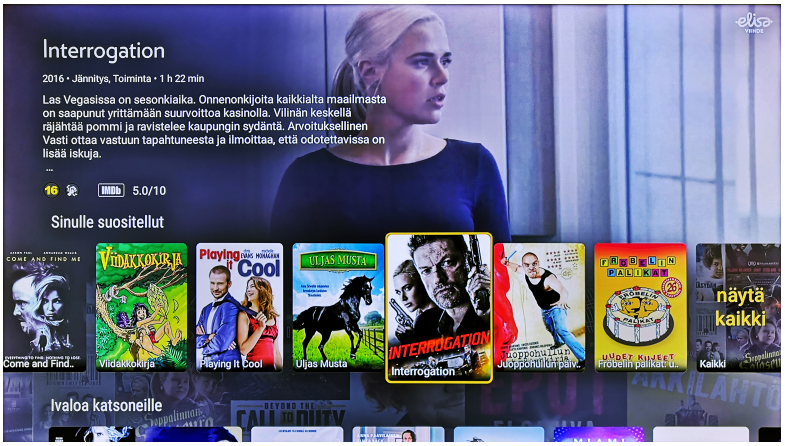


Figure 2.4: "Recommended for you" carousel in Aitio, Elisa Viihde (Android TV version).

demand content can be found in *Aitio* (en. "a box in a theatre"[10]), which is a subscription-based video library that contains both external films and series as well as Viihde's own productions. Users can also watch on-demand from *Vuokraamo* (en. "rental"[11]), where they can rent the content for a limited time or purchase it.

There are approximately 20 carousels on one page. Unlike Netflix, Viihde uses finite carousels, at the end of which there is a "Näytä Kaikki" (en. "Show All") button, which opens a grid with thumbnails showing possibly more videos belonging to that category. The number of videos within a carousel can vary on average from 10 to 40, depending on the device and there can be hundreds of videos in the "Show All" view. Figure 2.4 shows Aitio screen in Viihde's Android TV version. The layout is similar to Netflix's, with the description block at the top and at most two carousels displayed at a time at the bottom. The main difference is that the plot descriptions in Viihde are usually longer, there is always genre information, and instead of the match percentage, there are IMDb ratings provided for some videos.

In terms of recommendations, as of now, Elisa Viihde provides a personalised carousel called "Sinulle suositellut" (en. "Recommended for you"), which uses a cosine similarity to calculate both user and item-based recommendations (see Figure 2.4).

[10] https://en.wiktionary.org/wiki/aitio

[11] https://en.wiktionary.org/wiki/vuokraamo

2.5 Problem Reformulation

The size of TV screens and the way users interact with them – typically from a distance through a D-pad – impose constraints on how information should be delivered to the user. The layout should ideally be a one or two-dimensional grid with its elements organised so that there is always a straightforward and clear path between them, for easy navigation with directional keys. Additionally, the amount of text-based information shall be minimised.

In terms of recommender systems, the main algorithmic approaches are based on either rule, content or collaborative filtering, or their hybrid. Each approach has different strengths and pitfalls, hence the choice of the approach depends on the type of recommendations the system is aimed to provide, available data, their density and other resources, such as processing power and time.

Hence, the problem of efficient content discovery should be investigated from two perspectives: the delivery of content and related information through the user interface, and the generation of content suggestions. Furthermore, the implementation of the project this research was part of was focused on the front-end, thus there was an immediate need for UI designs. On the other hand, recommender systems mainly concern the back-end of a service. Therefore, the problem of content discovery in VoD services was divided into two sub-problems:

RQ1. How to present information on a TV interface to support efficient content discovery, given the limitations of the input device and the layout possibilities?

RQ2. What kind of recommender system approach could fulfil the needs users have in the context of content discovery?

and investigated in that order to follow the strict timeline of the project.

Chapter 3

Gathering Design Implications: Pre-Study

Watching TV is an activity not limited to any specific group of people as they can belong to any age group, have different background, skills or abilities. Therefore, it becomes a great challenge to develop an interactive system that can serve a diverse group of users. Designing for large populations is referred to as *universal usability* [31]. Following a user-centered approach to designing interactive systems (described in ISO 9241-210 [36] standard), involving users during the design process of interactive systems is crucial to improving its usability, acceptance and user satisfaction. Therefore, one of the first steps in this work was to conduct a user study with the aim to answer the following research questions (PS-RQs):

PS-RQ1. How users expect a VoD system to support content discovery?

PS-RQ2. What factors affect the choice of content to watch?

PS-RQ3. How well do existing services support content discovery considering those factors?

Answering above allows us to understand user needs and expectations towards VoD systems, what do they look for and how when searching for content to watch and, learning from the state-of-the-arts systems, identify opportunities for improving the existing content discovery solutions. Thus, the findings from the pre-study can be used to guide the design of the new service to answer RQ1, which impacts R2.

3.1 Method

The main methods for gathering data about a user are observation, interviews and questionnaires. Observations are great in uncovering the goals and tasks of the users and how tools they use support achieving those. With this method, we could investigate closely PS-RQ2, and partially PS-RQ1 and PS-RQ3. Observations can be conducted either in a controlled environment (laboratory) or in the natural setting (field). Each has its own strengths and weaknesses: in a laboratory, the investigator can control activities and tasks done by the study participant, which allows testing different conditions and easier identify problems, but is not suitable for learning about the context of use. Oppositely, observations in the field show how a system is used in the intended setting and the impact of the surrounding environment on it, hence producing more realistic data. The main disadvantage of studies in the field, however, is that they are time consuming and much more difficult to organise and conduct. [62, Chapter 7] While there are approaches to make field studies more "rapid", they typically involve having multiple investigators [49]. Therefore, a laboratory-based observation was the most suitable for two main reasons: both time and human resources for the user research part in this project were very limited (see Section 2.1).

To increase the reliability of the method and the certainty in the results it is recommended to *triangulate* by using different methods (*between-method triangulation*), measures (*within-method triangulation*) e.g. by involving multiple investigators (*facilitator triangulation*), or user groups (*user group triangulation*) [46]. Furthermore, since observations may not be able to uncover PS-RQ1 and PS-RQ3 exhaustively because they are not as explicit when using a service, it was necessary to triangulate the research approach. Hence, I decided to facilitate between-method triangulation and compliment observations with:

- **design brief assignment**, where the participants are asked to create a paper prototype of a VoD service, which would expose what they expect to have in a VoD system and and what do they find important in it (PS-RQ3), potentially uncovering the needs which are not yet satisfied by the existing solutions (PS-RQ2);
- **think-aloud protocol**, a method proved valuable in understanding users' thought processes in decision making [19, 60], and utilise it in both observations and design brief assignments to expose aspects that cannot be observed or deduced purely based on user's actions, thus providing more insight to all PS-RQs;

- **semi-structured interviews**, to explicitly ask about PS-RQ1 and PS-RQ3, also in terms of different use cases and contexts of use.

In terms of within-method triangulation, users from different groups were invited to the study to support user group triangulation and universal usability. Lastly, due to previously discussed constraints of the project, facilitator triangulation was not possible to incorporate in the pre-study.

3.2 Experimental Design

Within-subject design and *between-subject design* are the two main experimental designs. In the first case, all participants perform all conditions, while in the latter a participant and a condition are matched randomly. The advantage of between-subject design is that there are no effects of learning, while in the within-subject design the results of a condition may be affected by experience gained from the preceding one. However, to minimise the effect of an individual on a condition, a higher amount of participants is needed in between-subject design studies. [62, Chapter 14] Additionally, more statistically significant results can be obtained with less participants in within-subject design [33]. Hence, this experiment followed a within-subject design.

Tasks

Observations were carried out while participants were using Android TV versions of Netflix and Viihde. In each system, participants had to perform 7 tasks common for VoD services: finding a film and series they would like to watch, finding a specific film and series, finding a child-appropriate content, adding desired content to the favourites list and removing it. To further understand the users, while doing tasks participants were to *think-aloud*. It is a commonly utilised method in controlled laboratory settings that uncovers thought and cognitive processes of subjects while they are solving a task [19]. To minimise the learning effect of a task on a proceeding one, and the experience from using one service on the next one, both the order of tasks and services were randomised.

Apart from testing Viihde and Netflix, participants also had a design assignment. Participants were given verbal instructions shown in Table 3.1. To design a paper prototype, they were given: pens, papers, markers, sharpies, sticky notes, paper, scissors and pre-defined paper cut-outs of interactive elements commonly found in VoD services (e.g. buttons, drop-down lists,

Entertainment market The market and the need for online video streaming services are growing. There exist popular services like Netflix or Elisa Viihde, but there is potential to create a new service. **Team** Imagine that you are a designer who is a member of a team that decides to take on the challenge and design a new video service that could compete with the big players. **Assignment** Think what is really important in such a service, what would you love to have in it and what could make it stand out from the competitors. What do you think would be a perfect video service? Take some time to think and write down your ideas. Once you are done, design how the service could look like on the paper.

Table 3.1: Description of the design assignment.

carousels, video thumbnails), which they could freely modify. In this task, participants were asked to think-aloud during the process as well. Figure 3.1 shows a participant creating a prototype of a VoD service. The paper prototypes from all participants can be found in Appendix A.

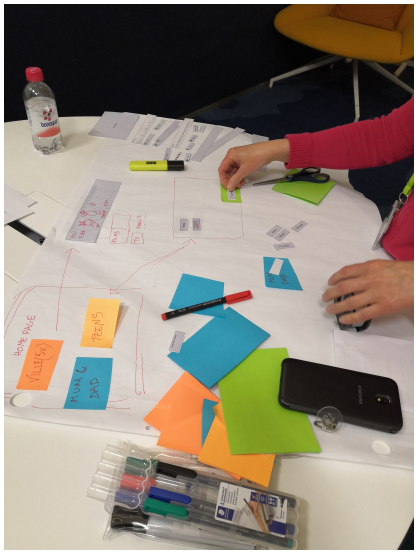


Figure 3.1: A participant creating a VoD service in the design assignment.

3.3 Participants

To support universal usability, the aim was to select participants representing different age groups and having varying prior experience with using VoD services. Two different invitations to participate in the study were sent to Elisa's customers via e-mail:

- Seeking *"entertainment watchers"* who regularly use VoD services like Netflix or Elisa Viihde, sent mainly to Elisa Viihde customers. To apply for participation, one had to respond with information about their age, examples of VoD services used, their frequency of usage and examples of content that they like to watch.
- Seeking people who are not familiar with VoD services such as Netflix or Elisa Viihde, sent mainly to customers who never had Elisa Viihde. To apply for participation, one had to respond with information about their age, whether they have tried VoD services like Netflix or Elisa Viihde and for how long.

Additionally, study participants were required to know both Finnish and English languages, since the study was conducted in English and one of the services was in Finnish only. Out of all submissions, 10 participants (6 females) in the age range between 23 and 57 (mean = 39, SD = 12.074) were selected. Table 3.2 summarises their familiarity with Netflix and Viihde.

Gender	**Age**	**Netflix**	**Viihde**
Female	26	+	-
	30	○	-
	32	+	○
	41	-	-
	51	+	○
	57	-	-
Male	23	+	+
	34	-	+
	41	-	-
	55	○	○

Table 3.2: Familiarity with Netflix and Elisa Viihde of pre-study participants. "+" indicates that a participant had used the service at least few times a week, "-" that it had never been used and "○" that it had been tested for a short period of time or less often than monthly.

5 participants did not have much prior experience with Netflix nor Viihde: 1 had a 1-month of Netflix trial multiple years before the study; 1 had tested Viihde and Netflix a few times; and 3 had never used either of the services. The remaining 5 participants had been using either of the services at least a few times a week.

All 5 entertainment watchers used VoD services on TV, 4 additionally on a mobile phone, 3 on a computer and 3 on a tablet. Every participant was compensated a 50€ gift card.

3.4 Apparatus

To run Android TV, a Technicolor Pearl B (4026) set-top-box was used. It was connected to a large screen 55" (1240.9mm x 770.8mm) Full HD TV from LG to create a living room experience. HUAWEI Honor 10 was used to: record video clips of actions on TV and think-aloud during observation, and record audio from the think-aloud process during the design assignments and interviews. Throughout observations, the camera was placed on a sofa table between the participants and the TV, ensuring that it is close enough to the participant to capture their speech clearly and that the image captures only the TV screen. Additionally, notes were taken during the experiment on a laptop.

3.5 Procedure

The experiments were carried out in one of the meeting rooms in Elisa's headquarters in Helsinki, Finland, which was styled to resemble a living room for the purpose of the study. The whole procedure took around 2 hours.

Firstly, participants were provided with snacks and drinks which they could consume during the experiment. This was done to create a more relaxed and home-like atmosphere. Before proceeding to the experiments, participants were given a consent and information confidentiality form. All participants agreed to take photos, video and audio recordings during the study for analysis purposes and non-commercial use. At the beginning of the study, participants were asked to imagine that they were at home and to sit comfortably on the sofa in front of the TV. Figure 3.2 shows one of the participants using Netflix in the experiment room. For tasks where participants had to find specific titles, they were asked to imagine that I was a friend who came over to their home and wanted to watch the given title. To find child-appropriate content, they were asked to imagine that there was a

4 year old child with us. At the end of each task, participants were asked to explain their choices. Additional questions were also asked if any action or part of the think-aloud was not clear. Each service was tested on average for 40 minutes.

Figure 3.2: A participant testing Netflix during the pre-study. The room was set to resemble a living room. The participant was provided with snacks and drinks, visible on the sofa table.

After testing both services, participants were introduced to the design assignment (see Table 3.1). They were given a few minutes to think and write down their ideas for a new service. Once a participant informed that they were finished, they were given an A2 paper, and other supporting tools for this task (see Section 3.2). The whole design task took on average 25 minutes.

The experiment ended with a semi-structured interview which lasted about 15 minutes. The aim was to investigate participant's video watching and browsing habits in different contexts. The questions were about:

- places and social situations in which they watch videos, whether these contexts affect what they choose to watch, and if yes – how and why;
- how do they tend to search for content to watch, how much time do they typically spend on it and why;

- what features or qualities do they usually look for in videos;
- their overall experience using VoD services.

In case of participants who had not used VoD services prior to the study, the questions were more general to watching videos, e.g. on live TV or in movie theatres, and their first impressions after testing the services. When formulating questions before and during the study, I was cautious to make them open-ended and not leading, e.g. *"Can you describe your overall experience using VoD?"* instead of *"Has your experience using VoD been good?"*.

3.6 Data Analysis

Verbal protocols from all tasks (using Netflix and Viihde, design assignment) and interviews were transcribed to a text format. Verbal protocols were then read carefully and converted to categories that described different themes and patterns [19]. A thematic and concept-driven approach was used in the study to avoid significant information loss, which is a result of using labels that describe atomic actions or thoughts [57]. It starts from creating an initial list of codes that represent concepts or more general ideas that are derived from previous work and hypotheses, and then expanding and modifying the list with concepts that appear in the transcripts [23]. New concepts from the think-aloud process were identified iteratively. Additionally, the analysis was complemented by re-watching video recordings of the interactions and creating codes for recurring problems that occurred.

I did the transcription and categorising process independently and fully manually.

3.7 Results and Design Implications

This section depicts the main results of the pre-study. The key findings from the pre-study related to content discovery are summarised in two tables: those that affect only the UI design in Table 3.3; and both UI and recommender system design in Table 3.4. In these tables, each key finding is followed by a design implication, which is then assigned a unique code used as a reference later in the text for simplicity. The results are further discussed in the context of pre-study's research questions in Sections 3.7.1 to 3.7.3. Finally, the Results are ended with a discussion of other issues identified during the study related to usability and user experience in Section 3.7.4.

Finding	Impact on Content Discovery	Design Implication	Code
Users rely on different metadata information (e.g. plot description, release year or country) when making a decision on what to watch, but it is not always provided or is not sufficient enough to make a decision.	Increased number of clicks and visited pages, if information is not provided in browsing views and the user needs to open a video page to obtain it.	Ensure that users can obtain metadata information with low interaction cost consistently whenever browsing videos.	MDB (Metadata: Browsing)
	Increased time and effort, if the user needs to use external services to obtain sufficient information.	Provide diverse information about a video.	MDD (Metadata: Diverse)
Users sometimes do not realise that they scroll content from the beginning in infinite carousels.	Increased: time and effort due to the necessity to browse the same content again; and risk of overchoice as the amount of videos appears to be higher.	Use finite carousels or provide a visual indicator that content is scrolled from the beginning otherwise.	CLI (Carousel: Indicator)
The number of carousels in a page is too high for most users.	Increased time and difficulty to find a carousel with a category of interest.	Allow easier scanning through categories.	CLS (Carousel: Scanning)
Search function was considered important by all participants, however it is not well supported in the studied VoD systems for discoverability, usability and understandability reasons.	Increased time and effort to find specific titles when the search function cannot be found, by having to browse through carousels instead.	Search function must be in a visible and easily accessible place in the system.	SVA (Search: Visibile & Accessibile)
	Increased time, difficulty and effort to find specific titles due to problems in using and understanding the search function.	Capabilities of search function and logic behind the search results, if more advanced than traditional title-based search, must be clearly communicated.	SCL (Search: Capability & Logic)
Users sometimes cannot find where they are currently on the screen and press random buttons to create movement and spot the focus point.	Increased time and effort if finding focus point is difficult and interrupts the browsing, decision making or thought process.	Make focus point easily recognisable and discoverable.	FPD (Focus Point: Discoverable)

Table 3.3: Key findings from the pre-study related to content discovery which affect the UI design.

Finding	Impact on Content Discovery	Design Implication	Code
Users tend to wonder why some content is recommended and do not trust recommendations that they cannot make sense of.	Increased difficulty to make a decision about items which cause confusion.	Provide an explanation why content is recommended.	REX (Recommend: Explain)
The number of carousels in a page is too high for most users.	Increased risk of overchoice.	Minimise the number of carousels as much as possible.	CLA (Carousel: Amount)
The same content is often displayed in multiple carousels or content that users already watched recently continues to show in carousels.	Increased: time and effort due to the necessity to browse the same content again; and risk of overchoice as the amount of videos appears to be higher.	Control and minimise the amount of duplicated content.	CTD (Content: Duplicates)
What content (carousels or videos within them) are displayed, and in which order may be different the next time the user visits a screen again.	Increased time and effort as users need to learn the placement of content on a page again; decreased discoverability as items found interesting previously may not be found again.	Avoid frequent and large changes in what content is displayed on a page and where.	CTC (Content: Changes)
Preference for content is complex and most commonly depends on the social context (who TV is watched with) and user's mood, thus may vary even within the same user.	Changing content discovery approach (e.g. which genre category is seeked) in different situations.	Provide content satisfying diverse range of preferences.	CTP (Content: Preference)
Being able to filter or narrow down content in a service is one of the most desirable features.	Faster content discovery if the volume of content is narrowed down based on the current preference.	Allow discovering content based on different features of videos.	CTF (Content: Filters)

Table 3.4: Key findings from the pre-study related to content discovery which affect both the UI and recommender system design.

3.7.1 Expectations Towards VoD Systems

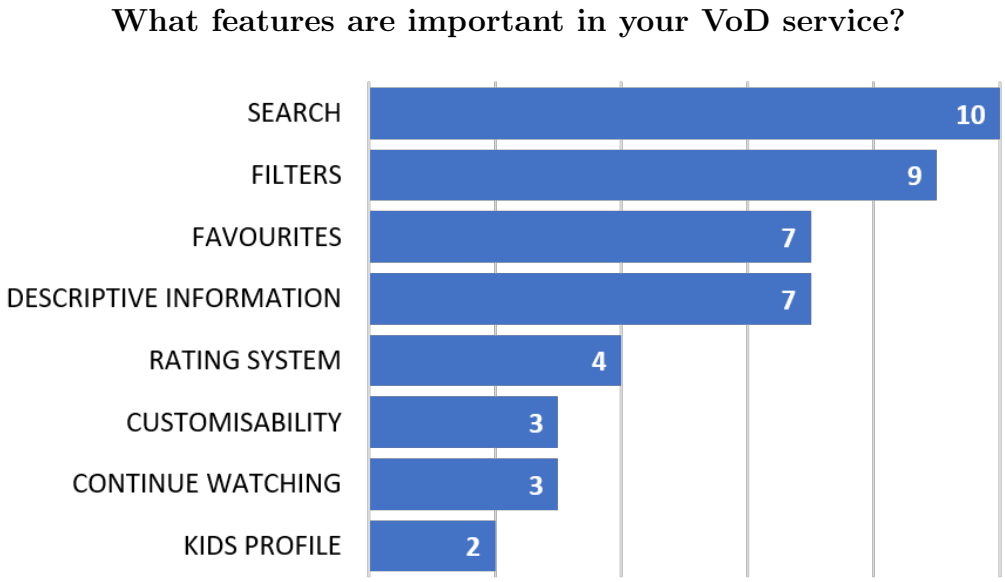


Figure 3.3: Features most commonly appearing in participants' paper prototypes made during the design assignment.

Design assignment required participants to create their ideal VoD service and hence exposed their expectations towards it (PS-RQ1). The features that participants commonly placed in their paper prototypes are shown in Figure 3.3. All of them paid attention to the fact that the search functionality needs to be in a visible spot in the service, which may be partially due to the difficulties participants had in using and finding it (see Section 3.7.2).

Moreover, most participants wanted to have the ability to narrow down the service's catalogue through different features of a video (filtering criteria, CTF). The most desired filters were: genre (9 participants), release year (6), external rating (e.g. from IMDb or Rotten Tomatoes) (5) and country (3). Other criteria included director, actor, feeling (e.g. happy or sad) and age restriction. 7 participants emphasised that in their ideal service, their own favourite content must be in a visible area. Among the remaining 3 participants, 2 did not have any carousels at all and their service contained only filters instead, giving them the full power over what subset of the content catalogue is displayed. 4 participants wanted to be able to rate the content and the main reason for that was to inform the system about the content they disliked and do not desire to see again in the future.

For most participants it was also important that they could see enough descriptive information (metadata) to make a decision on what to watch, which included plot descriptions, year of production, country of origin, language, or ratings. 3 of them expressed that in the existing services the support for metadata information was poor, which made it more difficult for them to make a choice (MDB and MDD, discussed further in Section 3.7.3).

3 participants wanted to be able to customise either the looks of the service or the carousels that are shown. Other 3 participants mentioned that "Continue watching" was the most important carousel and thus needs to be easily accessible, because they typically want to resume what they started watching in previous sessions. 2 participants wanted to have a separate "Kids" profile, which would not be shown in their own profiles nor affect the system's recommendations. Kids profile is currently not available in Viihde and was not found by any of the participants on Netflix.

To summarise, answering PS-RQ1, most findings highlight the fact that users expect to see only the content that may have interest in, and currently VoD services fail to filter out undesirable content appropriately, therefore making the users wish to have the ability to manipulate the content and service manually.

3.7.2 Preference for Content

To answer PS-RQ2, the most common factors that affected the preference for content to watch, as identified mostly from transcripts from observations and the closing interviews, are shown in Figure 3.4. 8 participants reported to typically watch TV with other members of their households and that they may choose content significantly different than if they were alone, e.g. *"with my wife then we watch some romantic movies. I personally like more some drama movies"*. From the remaining two participants, one was living alone and the latter claimed to have very similar taste to their spouse.

Participants also reported to commonly rely on recommendations from others, e.g. friends, family or critics and surprisingly only 3 were convinced by service recommendations. Common reasons to not choose from suggestions from the service was lack of trust in them caused by their ambiguity (REX), e.g. participants would start to wonder why some videos were recommended to them and question the quality of recommendations if they could not find a reason meaningful to them.

Furthermore, content preference depended also on the mood, as expressed by 6 participants. Moreover, there appeared to be a dependence between the mood and the depth of the storyline: 4 out of those 6 participants reported that their mood determines whether they would choose a video with

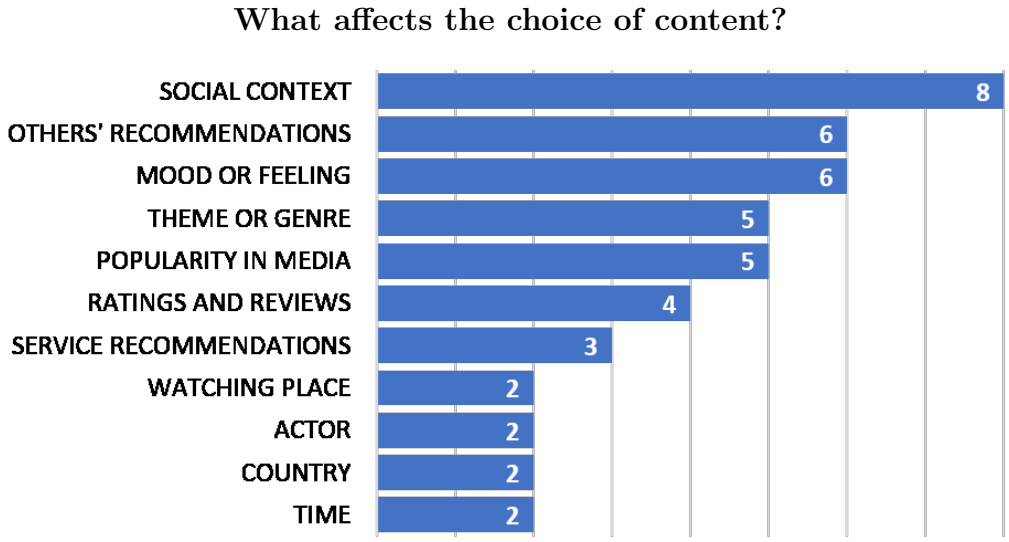


Figure 3.4: Interview result for the most common factors affecting the choice of content.

a lightweight plot, or a serious and deep one. For instance, one participant said that when in a *"bad mood or depressed, I might want something where I don't expect anything too serious (...) or otherwise like a movie which tries to make some statement, more serious movie"*. Additionally, the mood was affected by time as reported by some participants, who said that during different times of a day they are in different moods, which changes their preference for content to watch.

Since both systems currently offer mainly carousels with genre-based categories, an approach utilised by some participants was to try to match their mood to these categories. For instance, a participant expressed to feel like watching *"something lighthearted"* and started to wonder whether they could find it in romance or comedy carousels. This shows that offered categories are often not sufficient enough to support the different moods (CTP).

Additionally, the desired quality and length of a video seems to be affected by the place and time of a day: 2 participants reported that they leave content of higher quality to watch at the evenings or only from a TV, while when they are "on the go" and use a phone, they prefer to watch *"some random show"*. While this project focuses on Android TV, this finding is important for the future if the service is to be provided on other devices.

Most participants did not have any specific title or theme in mind when browsing the content. Instead, they tended to explore whether the system

has anything that could interest them. During observations, the content that was chosen by participants was dependent on different personal interests: each participant paid attention to different characteristics that the service displayed about a video, e.g. some were interested only in movies currently popular in the media, others in specific actors or countries. However, sometimes there was not enough information or it was not shown at all (MDD, discussed in more detail in the next subsection). The preference for content could also be guided by different motivations, e.g. self-development reasons instead of entertainment. For instance, one participant mentioned that while they usually watch American and domestic (Finnish) content, sometimes they would like to find French movies to *"brush up my French"*.

Findings for PS-RQ2 clearly indicate that preference for content is complex and often multi-dimensional, and may change even within the same user depending on the current situation they are in. Therefore, VoD systems should be able to satisfy their varying needs (CTP) to make content discovery more efficient.

3.7.3 Content Discovery in Existing Systems

Observations were invaluably helpful in identifying pain points in the existing solutions and therefore answering PS-RQ3. To summarise, the design of both examined services in some areas generated obstacles which prolonged the content discovery process and increased the risk of overchoice. Those obstacles include: increasing unnecessarily the amount of content to browse from (CLA, CLA, CTD, CLI); preventing the user from learning the content layout by changing it frequently (CTC); insufficient support for descriptive (metadata) information which is essential in decision making (MDD, MDD); poor usability and understandability of Search function (SVA, SCL); and interruptions during the discovery process caused by losing the focus point (FPD). These problems are discussed further in the following paragraphs.

High amount of content

Main difficulties were caused by the high amount of content to browse from, which elicited frustration or annoyance from 6 participants in Viihde and 5 in Netflix. First of all, finding a carousel with the category of interest was often time consuming (CLA) or impossible, e.g. some participants expressed to have spent more time than they would have liked to scroll through all the carousels to see whether there is anything that interests them (CLS).

Furthermore, observations and the design assignment exposed the following factors which additionally increase the amount of content to browse:

- Duplicates. In both services the same content often appeared in multiple services (CTD).
- Showing content that was already watched. 3 participants expressed annoyance that Netflix continues to show content in carousels despite them being already watched (CTD), e.g. one participant reported that *"shows that I watched and I wasn't able to remove them, and they still stay on top (...) it's annoying."*
- Infinite carousels. In Netflix the carousels are one-way infinite, meaning that once the user reaches the last video, the scrolling continues from the beginning. However, there is no clear visual indicator that this happens and one of the participants mentioned that *"sometimes I just don't notice that I'm going over them again"* (CLI).

All points above make content discovery process significantly more difficult and time consuming. Additionally, some participants expressed the desire to be able to remove from the Continue watching carousels content that they started watching, but disliked.

Frequently changing displayed content

In Netflix, the problem of inability to find the content of interest was also related to the fact that what is displayed may change from one visit to another. This includes: what categories are picked to be displayed in the carousels, what videos they contain and the order of either, all of which are decided by the recommender system (see Section 2.4.1).

During observations, 2 participants were unable to find the content (a specific series and category) that captured their attention in preceding tasks, because it was not displayed anymore where they remembered to see them previously (CTC). Both of them used Netflix daily and reported this to be a common problem. One participant mentioned that it was sometimes difficult to find even the "Continue watching" carousel, which usually is one of the first ones on the page, sharing during the design assignment that *"sometimes I even go to get something from the fridge and leave it for 5 minutes and it's already gone somewhere"*.

Poor support for metadata information

All participants relied on metadata information, such as plot description, release year or original country, when deciding what to watch. However, the support for descriptive metadata information flawed in two aspects: sufficiency to support making decisions and availability.

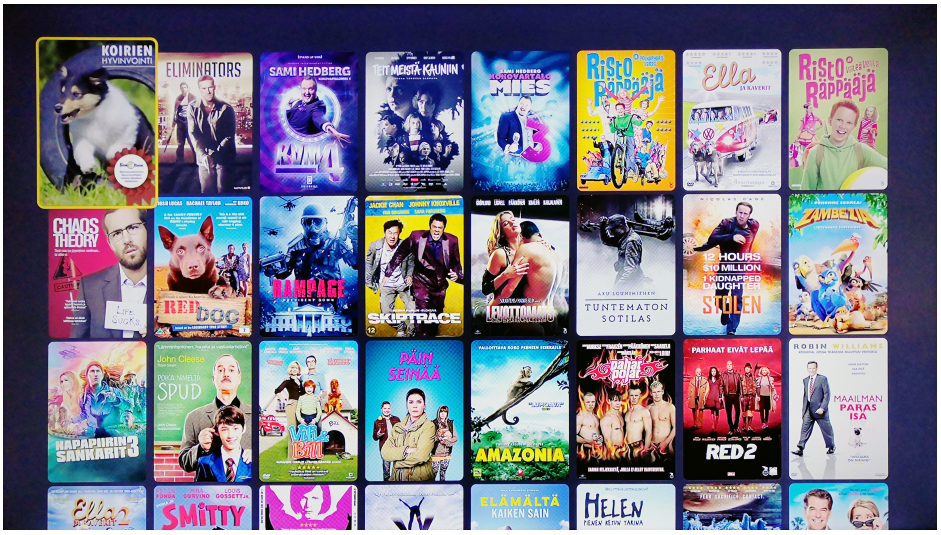
For instance, Netflix currently does not show how a video was rated by others and 7 participants reported that they usually check online ratings or reviews before watching a movie. 4 of them expressed that to watch mainly highly rated content and would in their natural setting check the rating from an external service (MDD).

Furthermore, in some views there were only video thumbnails without any descriptive information, e.g. in "Show all" view of any carousel in Viihde, "Netflix Originals" carousel in Netflix (see Figure 3.5), or search results in both services (see Figure 3.7). Therefore, users who were not familiar with titles or their thumbnails had to spend extra effort on opening video pages one by one to see the metadata information, which significantly slowed the discovery process (MDB). This issue caused one participant to give up and abandon the task of finding a specific title in Viihde, saying that *"this is again a mixed salad (...) You can go crazy. (...) You don't see the title and it would be nicer to read. Now I would have to go through all the videos. (...) I really don't want to."*.

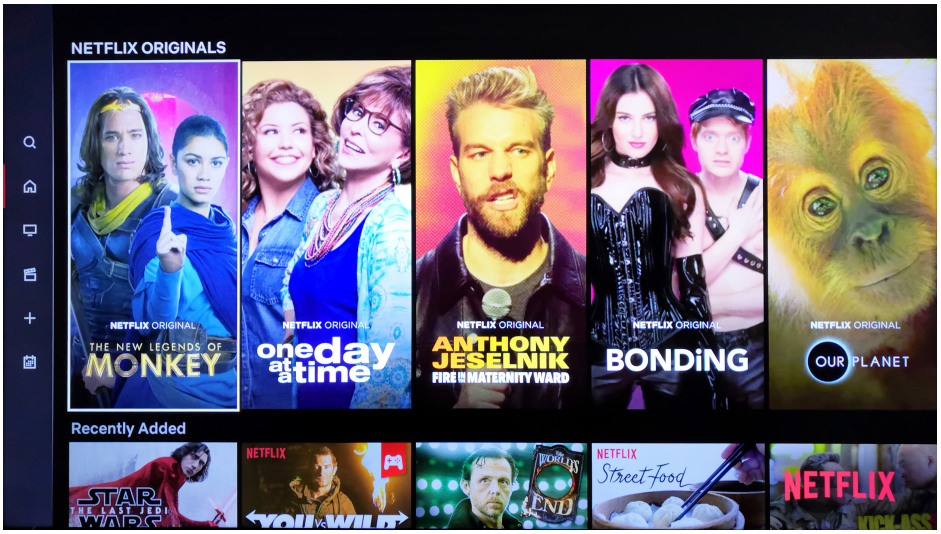
Lastly, in the views where series and movies are mixed together (e.g. "Home" screen in Netflix or some carousels in Viihde), it was difficult for participants to tell them apart while browsing, unless they were familiar with the titles (MDB). In the interview, 4 participants mentioned that they look for different features in series than movies, such as: release year (*"TV shows don't have to be super latest (...) 'Oh it's from 2011, I can watch this!'. Movies, if it's lately put there."*), what genre it belongs to or mood.

Discoverability of Search

Search is an important feature of a VoD system for many users (see Figure 3.3). It is the most straight-forward way to discover content when the user has something more specific in mind, which is often the case. For instance, as mentioned in Section 3.7.2, 6 participants reported to rely on recommendations from others and one of them shared that *"I usually don't start watching something if I don't know anything about it, so I'd have to hear about it from a friend."*. This shows that participants may come to a VoD service to search for titles they heard from others.



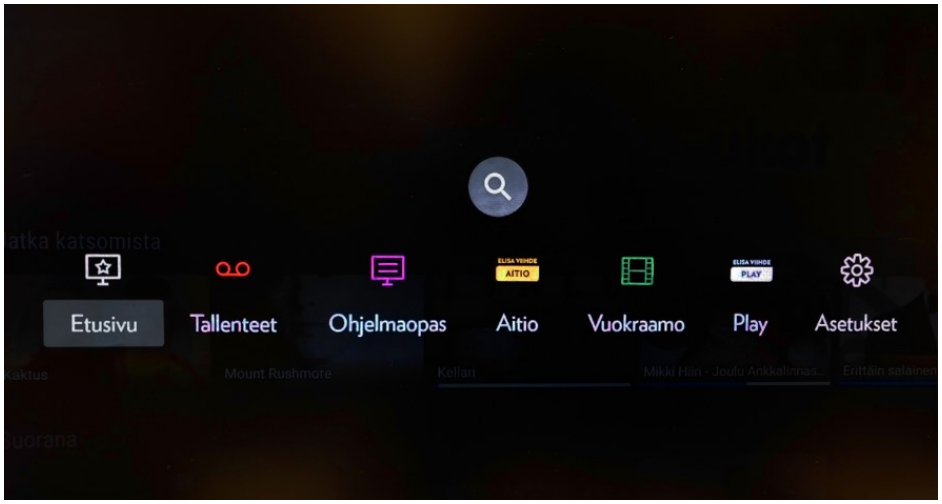
(a) "Show All" view of a carousel in Elisa Viihde



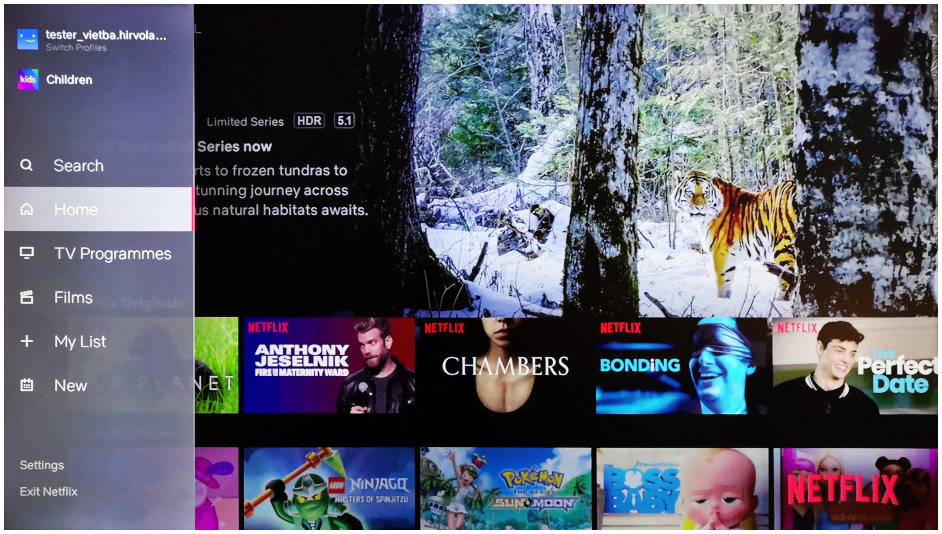
(b) "Netflix Originals" carousel in Netflix

Figure 3.5: Examples of browsing views with no descriptions and only thumbnails in Elisa Viihde and Netflix.

However, in Viihde 4 participants had difficulties to find the Search screen and other 3 participants never did (SVA). It is currently placed in the first row in the menu separately from other menu items (see Figure 3.6 (a)). There was a similar problem with Kids section in Netflix, where 4 participants who wanted to use it never found it either. Similarly to Search in Viihde, the



(a) Elisa Viihde



(b) Netflix

Figure 3.6: Menu in Viihde and Netflix

Kids menu item in Netflix is currently placed separately and above others (see Figure 3.6 (b)). The problems with visibility of these items may be potentially explained with Gestalt's principles of grouping, specifically the Law of Proximity and Similarity [81]: the users may have perceived those items as belonging to a different group of items and overlooked them, since they appeared too far or different than the rest of menu items. Another explanation in case of Viihde could be that the remaining menu items are

colourful in contrast to the search icon, which makes them more visually salient and captures attention away from the search icon.

Moreover, the design of the menu in Viihde may have affected the perceived effort needed to reach it when compared to Netflix, even though the number of clicks needed to enter the Search screen is the same or lower from any screen in Viihde. In both services the user needs to press: (1) "back" button to open the menu, (2) "up"-key to reach the search icon, and (3) "OK" to enter. In Netflix, the second step may need to be repeated depending on which menu item is currently selected. This paradox may be explained by identifying the main difference between the two menus: Viihde's menu has a semi-transparent dark overlay that occupies the whole screen, while in Netflix the menu overlays a narrow part on the left side of the screen. Therefore, the user experiences more screen changes, which may give an impression that the distance from the current screen to the Search screen is higher in Viihde.

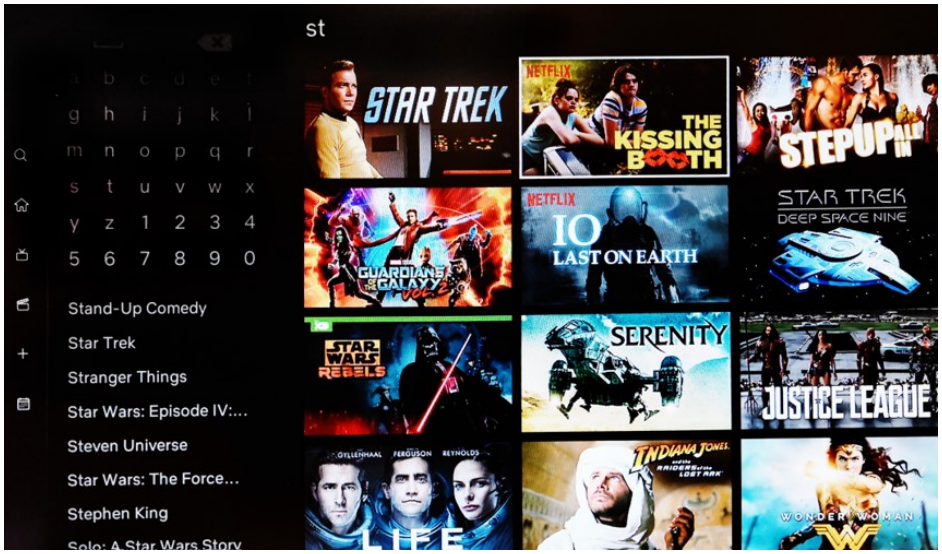
Understandability of Search

In Viihde's search, shown in Figure 3.7 (a), the main problem was with finding where the search results are. On the left side of the screen, there are multiple tabs representing different sources of videos and the search results are displayed inside them. The only visual cue that the results appeared in the tabs is the number of results next to the tab name. This kind of result displaying system does not match the user's goal for this task: to find a video matching the entered string. Instead, the service adds an additional step of choosing the source of the video. This was unexpected to all participants who tested it, as they were confused why the system did not display results matching the searched term, multiple even asked whether the task was tricky and the service, in fact, did not have the given title. This may have severe consequences since the user may be unable to discover content with the search function, even if it exists in the service.

In Netflix on the left side of search results the user can see a suggestion list, see Figure 3.7 (b). It contains both titles and more generic terms, which caused confusion. One participant thought it was illogical that it contains generic terms, because *"it indicates that there's a film called 'Stand-up comedies, but it isn't, because it indicates all stand-up comedies"*. Other 3 participants were confused about why the content of the suggestion list is different than that of search results. Suggestion list also displays titles that are not in the service, which annoyed a participant who saw a title they were looking for in the suggestions, but could not find it in the results after browsing through it.

Lastly, in case of both services, the search result currently displays titles which do not necessarily contain the searched string, e.g. "On Surfari" when the user typed "bullets" in Viihde or "The Kissing Booth" for "st" in Netflix (Figure 3.7). Therefore, it is clear from the observations that users' mental models of the search functionalities are significantly different than how they work in both services (SCL).

(a) Elisa Viihde



(b) Netflix

Figure 3.7: Search screen in Viihde and Netflix.

Losing focus point

On multiple occasions two participants seemed to lose track of where they are on the screen, sometimes asking me where the focus point was. To identify it, they would press random buttons on the remote to spot the movement on the screen. This issue may be related to age (both participants were over 50 years old) and potentially makes using the service difficult for a larger group of users.

While this problem occurred to only two participants and does not directly relate to content discovery, it does affect it, as the aforementioned actions interrupted and prolonged the browsing process (FPD), sometimes causing users to accidentally enter a different screen or leave the application.

3.7.4 Other Issues

The following is a list of other issues that were discovered during the study:

- The default Android TV keyboard used in Viihde, has a wide QWERTY layout. It is not optimal for typing with a remote, as pointed out by 4 participants. Keyboard with a similar number of columns and rows (e.g. Netflix uses a 6-by-6 square layout) could decrease the average number of key-presses between letters.

- To perform a search in Viihde, a confirmation key needs to be pressed, which frustrated two participants. While it is only one additional step, in practice it requires the user to click through all keys between the currently selected and the confirmation key.

- Fragmented service in Viihde caused problems not only with viewing search results, but also in:

  - Browsing content. The user needed to first decide which of the modules to enter (e.g. Aitio or Vuokraamo in Android TV) before being able to browse videos, and 9 participants did not know how their content catalogues differed. This was a problem also for Viihde users, e.g. one participant has *"been having Elisa Viihde for many years in my home, but I've never been using Aitio. I think this is the place where there is Elisa Viihde own things"*.
  - Finding favourited videos. Each module has its own Favourites lists. 4 participants expected to find them in Tallenteet (en. "recordings", the place where recordings from live TV are stored). 2 out of 3 participants who added a video to favourites from

Vuokraamo spent over 10 minutes to find it, because they assumed that a video which was already added could not be in rentals.

- Reading title from a thumbnail may be difficult considering its small size, or impossible if the thumbnail does not show the title. This was a problem in Netflix and "Show all" views in Viihde, where titles were not displayed in a text form near the thumbnails.

- Rewind and fast-forward in the video player used in Viihde can be currently done in steps of 30 seconds or 5 minutes. This, however, does not support the user's goal in these tasks, which is to find a scene of interests. Typically when watching a video, users are not aware of how far in terms of time a particular scene is. Instead, determining the scene of interest usually is based on the visuals of it.

Chapter 4

Interface Design Proposal

After the user research phase, the next goal was to translate findings from the pre-study into design decisions and wireframes that demonstrate *how to present information on a TV interface to support efficient content discovery* (RQ1). Thus, this chapter depicts the design decisions and resulting wireframes proposed based on the implications identified in the previous chapter. The wireframes were then turned into polished interface designs by the visual designer, which were delivered to the developers. During the project, I worked closely with the visual designer and developers to ensure that the principles behind the new interface are understood by team members and implemented accordingly.

Due to the time constraints of the project, the team decided that in order to ensure a better quality of the service the development should focus on the re-implementation of Home screen's interface, which is the main screen where a user can browse content. Therefore, more efficient discovery of content had to be achieved only within that one screen. Table 4.1 summarises the decisions regarding the interface design in relation to the design implications, and a new browsing approach introduced in the Home screen using a carousel with categories, referred to as a *Category carousel*. The main components of the Home screen are: menu, description block, carousels and the Category carousel (see Figure 4.1). The design and wireframes of each are described further in the following sections.

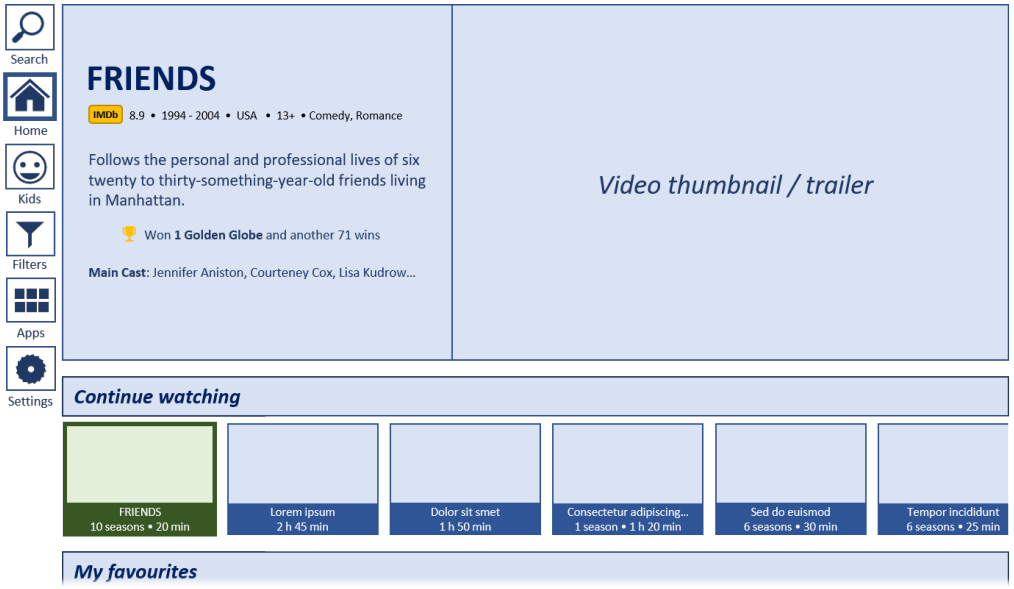
4.1 Menu

On the left side of the screen, there is a narrow icon-based menu which is always visible, regardless of the screen. Therefore, even if the user browses through menu items, they in practice are still in the current screen, mak-

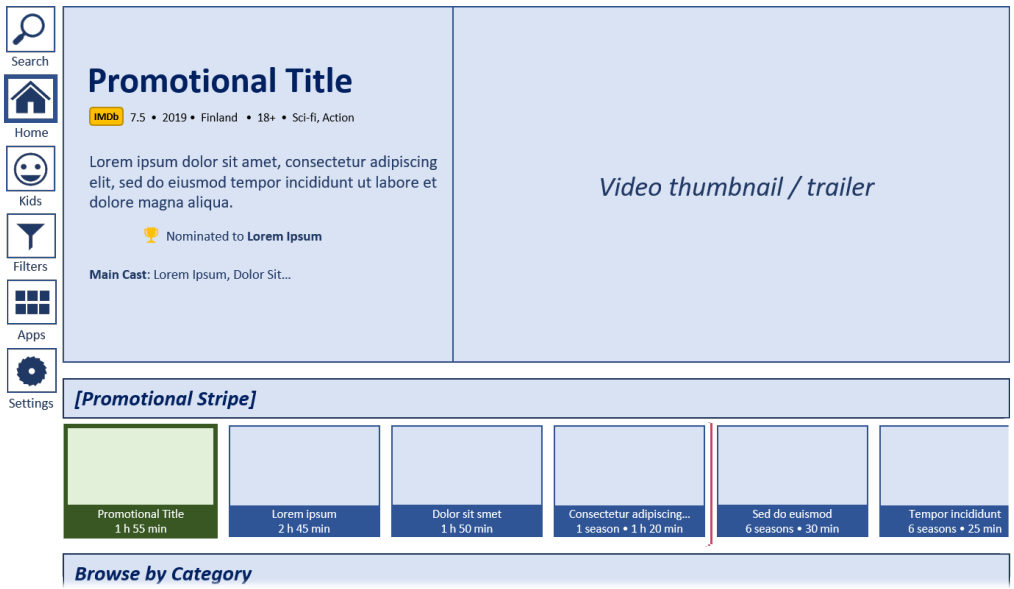
Code	UI Design Proposal
MDB (Metadata: Browsing)	Always provide a description block with metadata information about the currently selected video in all views where the user can browse videos. The description block is visible without the need to expand or open another page.
MDD (Metadata: Diverse)	Include in the description block different key metadata information which are commonly important for users when comparing videos to watch (as identified in the pre-study). Display them in a consistent layout and format throughout the service.
FPD (Focus Point: Discoverabile)	Highlight the focus point with a border using a prominent colour. Keep the focus point in all carousels on the left most video tile (requires using a one-way infinite to the right carousel).
CLI (Carousel: Indicator)	Display a separator between the last and first video in a carousel.
SVA (Search: Visibile & Accessibile)	Place search as the first menu item and display it in the same hierarchy as other items. Always display the menu on the left side of the screen.
REX (Recommend: Explain)	Provide a brief explanation near the description block of why the currently selected video is considered relevant to the user, for every video which is displayed as a result of a recommender system.
CLA (Carousel: Amount)	Limit the amount of carousels to the categories that always need to be in the service and allow browsing other categories through a carousel (*Category carousel*).
CLS (Carousel: Scanning)	Display multiple categories in one row on screen with the *Category carousel*.
CTP (Content: Preference)	Offer in the *Category carousel* a diverse choice of video features that users may be looking for in different situations, e.g.: depth of storyline, rating, type of characters, genre and so on. The category can represent a single or compound features.
CTF (Content: Filters)	Allow browsing by different features of videos with a *Category carousel*, where categories are essentially different filtering criteria.

Table 4.1: Proposed UI design decisions based on design implications identified in the pre-study.

ing screen changing perceived as taking less effort in contrast to full-screen menus. To support SVA, the Search function was placed on top of the menu.



(a) First static carousel



(b) Last static carousel

Figure 4.1: Wireframes of the Home screen in the initial view with the first carousel and after scrolling down to the second last carousel.

Moreover, following Gestalt principles, the Search menu item is the same size and distance as other menu items, and its icon uses the same colour scheme. These decisions ensure that the Search menu item is perceived to have the same importance (belonging to the same group) as the remaining menu items and can be easily spotted by the user, which was a major issue in Viihde (see Section 3.7.3, Discoverability of Search). Since Kids section in Netflix also had poor discoverability, it was placed in the menu with similar principles as Search.

4.2 Carousels

To minimise the number of carousels (CLA) I proposed to divide the categories into static and dynamic. Static categories are those, which need to always be displayed in the service, while dynamic ones are those that adapt to the user over time. Static categories are displayed in separate carousels as first in the Home page and are advised to be in limited quantity (see Section 4.2.1). Dynamic carousels are shown through a 2-level Category carousel, discussed later in Section 4.2.2.

To satisfy FPD, the navigation is designed so that the focus point in carousels is always on one side of the screen, allowing the user to learn its position on the screen and hence minimising the risk of it being "lost" while browsing. With this type of interaction, to avoid showing empty carousels, it is required that the carousel is one-way infinite. This means that it wraps its contents on one end and can be thus scrolled infinitely in one direction. In this case, the carousel is one-way infinite to the right, allowing the focus point to always stay on the leftmost on-screen video in the currently browsed carousel, while the contents are animated to move towards it as the user scrolls (e.g. leftwards if the user presses the right key). Since the carousel is one-way infinite, to ensure that the user can identify that they are starting to scroll from the beginning, there is a separator placed between the last and first item (CLI), which is shown in Figure 4.1 (b) between the fourth and fifth video in the carousel dedicated for promotional content.

To support MDB and MDD in relation to the issue of distinguishing movies from series, under each video thumbnail there is a label displaying the title and either the duration in case of a movie or the number of seasons with an average duration of an episode for series. Additionally, with this solution, the user does not have to read titles from thumbnails, which was reported to be problematic in the pre-study. Movies and series can be quickly differentiated from the duration and season information, and videos which exceed the time a user can allocate on watching TV can be easily skipped.

4.2.1 Static Carousels

Static carousels are limited to: Continue watching and My favourites, as they are often visited when a user re-enters the service as identified in the pre-study; and operator-enforced carousels required for business reasons, such as one containing only original productions (CLA). While in theory operators have the freedom to define many carousels, is it advisable to keep the amount low (e.g. no more than 3) and allow the user to focus on the personalised content.

Moreover, Continue watching is prioritised over My favourites and operator-enforced promotional carousels, since users reported to usually continue with where they left off the service in previous sessions. Figure 4.1 (a) shows the initial view of the Home screen, where Continue watching is always the first carousel, and (b) shows the state when the last static carousel is selected.

To support MDB, the description block for the currently selected video is always visible above the carousel it belongs to.

4.2.2 Category Carousel

Research has shown that grouping information into higher-level categories can aid scanning and understanding it [22, 48]. Hence, in information architecture, it became a good practice to structure information into chunks, which are commonly distinguished with labels or separated with white spaces or colours [67], following Gestalt's Law of Proximity [81]. However, currently in both Viihde and Netflix, users can see at most two carousels at a time on screen, forcing them to scroll down a long page to discover what categories are available, thus increasing the interaction cost of scanning through available categories. To solve this problem, instead of having a long list of carousels displayed sequentially, a single carousel that contains all available dynamic categories was introduced (CLA). Shown in Figure 4.2, *Category carousel* is a 2-level carousel which scrolls up the screen when the user enters the first category in it.

When the user selects a category from it, the carousel below it is populated with videos belonging to that category. The second level is essentially a carousel with content filtered by the selected category, hence giving the user the ability to quickly narrow down the content to the one belonging to the selected category (CTF). This level is referred later as *Filtered carousel*. Therefore, the Category carousel displays content in a hierarchical manner: it shows the parent nodes – categories – in the first level and the child nodes of the selected parent in the Filtered carousel below it.

Figure 4.2: Wireframe of a view with the Category carousel ("Browse by Category"). The bottom carousel, called Filtered carousel, displays videos in the category currently selected in the Category carousel.

According to the guide book on information architecture:

> *"We have organized information into hierarchies since the beginning of time. Family trees are hierarchical. (...) Because hierarchies provide a simple and familiar way to organize information, they are usually a good place to start the information architecture process. The top-down approach allows you to quickly get a handle on the scope of the web site without going through an extensive content inventory process."* Rosenfeld et al. (2002), p. 33 [67]

Hence, the main benefit of a Category carousel is that it allows the user to scan through more categories within one screen using a top-down hierarchical view – a structure familiar to people (CLS). This is faster than having two categories additionally separated by video carousels visible at a time, on a long page with sequentially displayed content.

When building hierarchies it is also important to consider their breadth and depth [67]. With a deeper hierarchy, there would be fewer categories in one level and reaching the content would require going through more levels (sub-categories). Therefore, in the context of TV interfaces it was more suitable to have a broad, but shallow 2-level hierarchy, because a more complex hierarchical view would be notably harder to navigate with a remote control (see Section 2.2).

An important idea of Category carousels was also to have it include categories representing different kinds of features that characterise a video, such as the depth of storyline or its rating, which were identified as important factors affecting the preference for content in Section 3.7.2. Moreover, the categories could represent one (e.g. "Light Storyline") or compound features (e.g. "Most Watched in 2019"). Providing categories representing different features of a video in the Category carousel allows fast discovery of diverse type of content, which is needed considering the changing nature of content preference (CTP). What categories could be initially shown in the service and how they could be adapted to the user are discussed in Section 5.3.1.

Lastly, to support MDB, the description block is shown in the Filtered carousel for the currently selected video together with it in a larger tile than others.

4.3 Description Block

As mentioned in the previous section, MDB is satisfied by always showing a description block, either above static carousels or inside the Filtered carousel in the Category carousel view. To support MDD it was essential to decide on which information to show while keeping in mind that it is not recommended to display a lot of textual data on a TV interface (see Section 2.2).

Newell and Simon's (1972) [57] research on problem-solving suggests that in order to be able to process information within the bounds of one's limited capacity, people utilise heuristics to reduce the cognitive strain. Payne (1976) [60] discovered that people tend to check the same parts of information when presented with two alternatives, however in case of multi-alternative problems the amount of processed information is reduced and which parts of it are processed between options is not consistent. Therefore, it was suggested to provide information about options which make them easily distinguishable. Furthermore, according to Gigerenzer et al. (1999) [24], providing less information with only key points about an item is in fact more insightful, as humans tend to focus only on a small amount of information that is important to them, ignore the rest and use simple heuristics based on those information cues to make a decision.

Such behaviour could also be seen during the observations in the pre-study, as different participants seemed to pay attention to different aspects of a video, depending on their preferences and motivations (see Section 3.7.2). Hence, to support efficient decision making, the key information that was selected for the description block was based on what participants commonly reported to take into consideration when choosing content to watch.

Moreover, how information is displayed on the screen is also crucial, as it affects how easily a user can spot information that is relevant for them to make a decision. Creating chunks of information aids that by making scanning through them and finding the relevant bits require less effort, which is especially crucial when the user is under information overload [22]. Following this, information about a video was divided into four groups separated with spaces between them, as can be seen in Figures 4.1 and 4.2. The following are metadata information that can be found in a description block, displayed consistently in the following order from top to bottom:

1. Short characteristics of a video: rating, release year, country, age restriction and genres.

2. Plot description. Since the storyline is usually one of the most important factors affecting the choice of content (see Section 3.7.2), it is further emphasised and distinguished from the remaining information by using a larger font size, which also makes the text more readable on TV, as it is the longest chunk of text in the description block.

3. Awards, if there are any. While the importance of awards was not mentioned during the pre-study, they are achievements which can make a video stand out from other options and therefore making it a piece of salient information about the item. Therefore, to increase its visual saliency and thus allow the user to easily spot an awarded video even without reading the description block, an icon with a bright cup is always placed next to the award information.

4. List of the main cast, advised to be limited to top three for clarity.

5. Explanation of why a video is considered relevant to the user. This information is shown in the Filtered carousel, because unlike static carousels, it is not directly manipulated by the user and personalised to the user with a recommender system. Therefore, to satisfy REX, the description blocks for videos that were selected based on user's past behaviour must always include an explanation. In Figure 4.2, the explanation is displayed below the video's thumbnail and descriptions.

Moreover, Preece et al. (2015) [62, Chapter 3] suggest that more information should be easily accessible for users who want to know more about an item. Therefore, more information, such as the directors can be shown if the user is intrigued enough to enter the video page.

Lastly, it is important to note that while for this project the development was focused on the Home screen, a similar approach of displaying information was suggested for the Search screen's results to satisfy MDB.

Chapter 5

Recommender System

In the previous chapters I identified design implications for the UI design in the context of content discovery based on findings from the pre-study and background research on designing for TV. Some of the implications also affect the design of a recommender system and constrain the approaches that could be potentially used when building it. For instance, the need to explain recommendations limits what types of algorithms could be applied, and the category-based browsing could benefit from personalising the selection of categories that are displayed to the user.

While the implementation of a recommender system was not in the scope of the proof-of-concept project, discovering *what kind of recommender system approach could fulfil the needs users have in the context of content discovery* (RQ2) was important for the potential future development of it. Furthermore, having a good recommender system is important for improving the content discovery process and avoid the problem of overchoice.

This chapter depicts steps taken to answer RQ2 and proposes design characteristics of such a recommender system and data it would need. Table 5.1 summarises the proposals made for building a recommender system, which are discussed in more detail in this chapter, as follows. Firstly, possible solutions to the design implications that resulted from the pre-study are investigated in the existing literature in Section 5.1. Secondly, the data that is available in Elisa Viihde is used as an example of real-world metadata that could be provided by a potential operator, and their sufficiency for a VoD recommender system are discussed in Section 5.2. Lastly, more concrete design and implementation suggestions for a recommender system are presented in Section 5.3.

Code	Recommender System Design Proposal
REX (Recommend: Explain)	Use an interpretable model that recommends items based on their features.
CLA (Carousel: Amount)	Select categories for the Category carousel based on what categories and type of videos a user interacted with (*category recommendations*).
CTF (Content: Filters)	Select videos for categories in the Category carousel based on what videos a user interacted with *within* each category (*within-category recommendations*).
CTP (Content: Preference)	Keep balance between relevancy to the user and diversity in both category and within-category recommendations and use time dimension to account for different contexts.
CTD (Content: Duplicates	Identify which categories are often visited by the user and ensure that no duplicates are displayed in them. Remove recently watched videos from recommendations.
CTC (Content: Changes	Adapt the frequency of changes to how active a user is and introduce the changes gradually.
IFB (Implicit Feedback & Balance)	Collect behavioural data for modelling implicit feedback and generate implications of both positive and negative feedback to keep a balance between them.

Table 5.1: Proposals for the recommender system based on the pre-study, UI design decisions and literature review.

5.1 Literature Review

To further investigate design possibilities for a recommender system, I did a literature review in terms of explanations (REX), utilisation of contextual and mood information (CTP) and needed user feedback to fine-tune the system. The following are the main suggestions for recommender systems in the context of this project made by comparing arguments for and against different approaches found in the literature:

- REX (1/2): Studies on explanation types showed that explanations referencing previously consumed items and features they have in common with the recommended item are the most helpful for making de-

cisions. Therefore, *explanations should be based on feature-item references* (FIX: Feature-Item Explanation).

- REX (2/2): Empirical research showed that high prediction accuracy is not necessarily correlated with user satisfaction and moreover, explanations improve the reliability and usefulness of recommendations. Additionally, recent regulations require algorithms to be transparent. Thus, *the system should use interpretable models*, rather than black-box ones (INM: Interpretable model).

- CTP: Context of use may change user's preference for content, therefore it should be considered by the recommender system. Currently, most solutions in research that utilise contextual information beyond the time dimension require the user to input it manually or would need obtrusive measuring or tracking devices. Hence, the *time dimension is still the most feasible information used to differentiate between different contexts of use and should be utilised in the system* (TCX: Time context).

- Additionally, the topic of user feedback, which is needed to train recommender systems, was explored in this section. Implicit feedback appeared to be more suitable than explicit one from the user experience perspective, and it also allows to generate significantly more data. However, implicit feedback is typically built around the indicators of positive feedback, e.g. if a video is watched it is assumed to be liked by the user. Therefore, the *system should model implicit feedback and collect more indicators of negative feedback*, to counter the problem of imbalance between the indicators of positive and negative feedback (IFB: Implicit feedback and balance).

The following subsections discuss in more detail the arguments found in the literature.

5.1.1 Explanation of Recommendations

During the observations in the pre-study (see Section 3.7.2) participants often wondered why a video was recommended to them and wanted to understand the rationale behind it (REX). Previous research suggests that users want to understand why items were recommended to them [30, 84] and explanations have been shown to have a significant impact on user acceptance and satisfaction of intelligent systems [8].

It is a part of human intelligence to seek rationale behind one's decisions or beliefs and infer them, e.g. by guessing others' intentions in social interactions [69], and the ability to do so is called the *theory of mind* [63]. Hence, when faced with decisions or beliefs that are unexpected, people tend to ask "Why?", which was observed in the pre-study when participants wondered about the possible rationale behind an unexpected for them recommendation (see Section 3.7.2). The reason why this behaviour happens not only between humans, but also in the interactions with intelligent systems, is that users unconsciously assign human-like characteristics to them and therefore expect in response explanations from the system [50].

There are many benefits of explaining recommendations from the user experience perspective and the most important include:

- Increased trust in the recommendations and their perceived reliability [8, 27, 30, 34, 50, 69];
- More awareness of how one's actions affect the underlying algorithm as a result of forming an accurate mental model of the recommender system, and consequently, the feeling of control and ability to adjust recommendations to one's advantage if desired [1, 27, 50];
- Better understanding of the strengths and weaknesses of the system, and more understanding towards inaccurate recommendations if the provided rationale seems sensible [27, 30, 69], e.g. if recommendations are related to videos watched with a friend that would not be otherwise chosen;
- Improved decision making, as the user is provided more information to assess a video and can decide how much confidence can be put in the recommendation [27, 30, 84], e.g. if the suggested video is explained to be similar to one they are already familiar with, they can better determine whether it fits their mood.

Types of Explanations and Assessment Criteria

Moreover, the suitability and type of explanations for recommender systems need to be evaluated [30]. Vig et al. (2009) [84] collected from previous research three main assessment criteria:

- *Justification*, the ability to support the user's understanding of why a recommendation was made;

- *Effectiveness*, the ability to aid the user in decision making;
- *Mood compatibility*, the ability to help the user identify whether the suggested item matches their current mood.

Research on explanations uncovered that statistical measures showing the system's confidence in the recommendation are not important to an average user and instead, what caused the system to recommend the item is more insightful [50]. Explanations can be divided into different types based on "resources" used to form them: user, feature or item, as illustrated in Figure 5.1 [84]. Feature-based explanations (e.g. referencing an actor or genre) has been shown to be superior to user-based and item-based ones for recommender systems, because the connection between two items may not be obvious to the user and similarity to another user is too ambiguous to understand by an average user [7, 30, 34, 84]. Furthermore, Papadimitriou et al. (2012) [59] investigated different combinations of those resources in terms of the above-listed assessment criteria and discovered that feature-item hybrid (e.g. $item\,X$ was recommended, because it contains $feature\,A$ which is present in previously liked $item\,Y$) was most favoured by users and obtained highest satisfaction ratings in all three criteria.

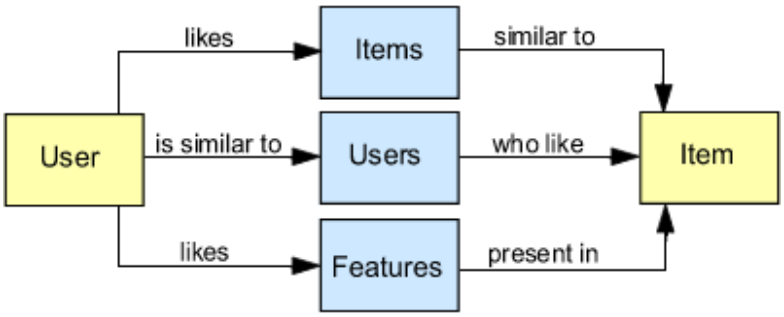


Figure 5.1: Resources used for explaining the relationship between the user and the recommended item (Source: Vig et al. (2009) [84]).

Hence, it is important to know not only what items previously interacted with are similar to the suggested one, but also *how*. For instance, in terms of VoD services, the system could recommend a movie, because multiple of the previously watched movies or series (items) belong to the same genre (feature) and refer to those items and features in the explanation (FIX). Additionally, the way explanations are worded and what kind of features are used for justification are also important. Vig et al. (2009) [84] discovered that users prefer when explanations contain facts rather than subjective terms

(e.g. "violence" instead of "violent"), and general themes (e.g. genre) over specific topics (e.g. amnesia).

Algorithmic Approaches in Relation to Explanations

The necessity to provide meaningful explanations to the user (REX) enforces that the algorithms used in a recommender system should be explainable and based on features of items. Following that, rule-based approaches are simplest to explain, and CBF methods using item's features are more suitable for discovering items similar to those interacted with and features they share [1, 59].

Despite of that, recommender systems are nowadays often implemented with so called *black-box* algorithms, which are highly non-linear approaches that are opaque in terms of how they arrived at their predictions, therefore making the rationale behind their outputs unknown even for their developers [30, 69]. Their popularity is explained by improved performance for solving complex tasks in comparison to linear models. Figure 5.2 [27] compares different algorithms in terms of their performance and explainability.

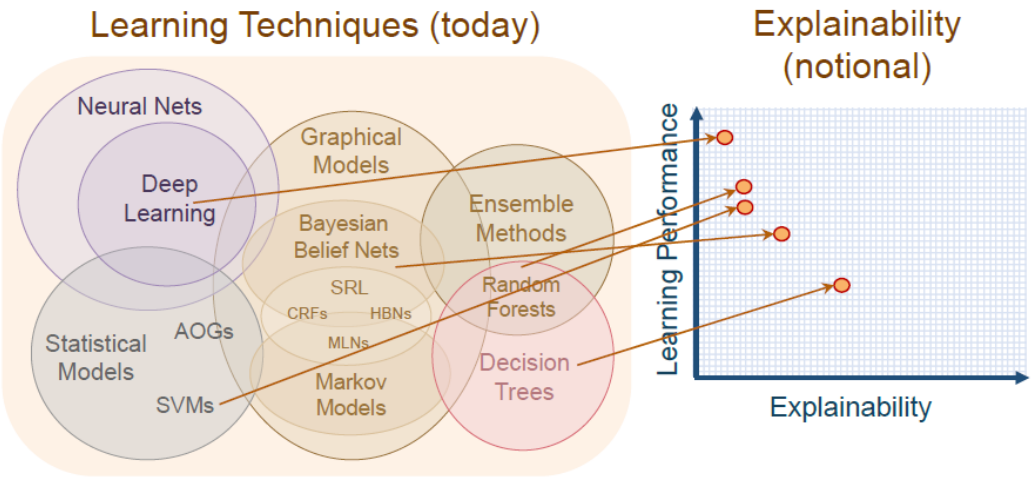


Figure 5.2: Comparison of learning algorithms in terms of their performance and explainability (Source: Gunning (2017) [27]).

In order to maintain the benefit of black-box algorithms' high performance while meeting the need to explain their predictions, rapidly growing efforts in the area of explainable artificial intelligence (XAI) focus on creating machine learning (ML) techniques whose purpose is to "open" the black-box

and find explanations to what the model has learnt or its individual predictions [8, 27, 69]. Using an algorithm to explain another AI algorithm has shown promising results, e.g in image recognition [69], however, may not be a feasible solution for multiple reasons.

First of all, as noted by Abdul et al. (2018) [1], much of current XAI research focuses on algorithmic approaches to generating justifications, often not guided by knowledge from cognitive psychology nor evaluated with real users, rather than their usability and effectiveness for the user. Moreover, justifications are not equivalent to transparency, since they do not necessarily reveal the true mechanics of the algorithm [84]. Transparency became especially crucial since the introduction of recent General Data Protection Regulation (GDPR) in the European Union[1], which gives users of intelligent systems the *right to explanation* for how the algorithm came to its output; and the joint statement from ACM's US Public Policy Council and Europe Council regarding algorithmic transparency and accountability[2] [1].

Therefore, transparency may not be completely ensured with the aforementioned approach proposed in XAI, since justifications are essentially based on another algorithm's guess, which may, in fact, be misleading and misinforming the user and become subject to legal liability. For this reason, a secure alternative to black-box models is still using interpretable models (INM), such as linear classifiers and rule-based models like decisions trees [1, 8]. Moreover, while high accuracy of predictions is an *objective* measure fundamental from the algorithmic perspective, research on the user experience and usefulness of recommender systems showed that it is only a constituent of the *perceived* accuracy. In fact, users may prefer recommendations generated by a system with worse prediction accuracy in favour of better explainability and transparency, but also other aspects, such as the ability to suggest diverse content and consider different contexts of use [12, 40, 64].

5.1.2 Context and Mood Awareness

As identified in Section 3.7.2, the context in which TV is watched and the user's mood have a great impact on the perceived relevance of the video to the current situation (CTP). In research on recommender systems, the importance of context led to the development of *context-aware* approaches, shown to improve the accuracy of recommendations by modifying existing methods, such as CF or CBF, with an additional context dimension [28].

[1]European Union General Data Protection Regulation https://eugdpr.org/

[2]Statement on Algorithmic Transparency and Accountability by ACM U.S Public Policy Council and ACM Europe Policy Committee. https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf

The most often used dimension is time, as it allows to uncover temporal changes in user behaviour, e.g. caused by period-dependent consumption on holidays, without requiring the input from the user [28, 73]. To simplify the time dimension, it is commonly divided into discrete categories, such as workdays and weekends [28, 45], days of the week [16], day and night [45], or more granular ones like 30 minute time slots [58].

Obtaining other dimensions, such as social context or mood, however, are more challenging. Automatically inferring users' mood or who they are with would require more advanced technology, e.g. using image-based recognition [43, 51, 89] or physiological measures [38, 82, 87]. However, these methods are too invasive and obtrusive, especially for an in-home leisure activity.

A common approach in context-aware and mood-aware systems is to require the user to manually input information about their social context [16] or mood [53, 73]. This solution is not ideal either, as in the context of TV interfaces, providing such input may be a tedious task with a remote control or users may be unwilling to go through additional steps in their leisure time. Wang et al. (2010) [85] used the mood of movies from the metadata, with the assumption that they reflect user's mood, however, they found that including mood information did not improve significantly the predictions. One explanation for that may lay in the assumption itself. While for some people preferred video's mood may be the reflection of their own, the opposite relation was identified in the pre-study, where some participants reported to seek lighthearted content when they are in a bad mood, most likely as a way to repair it.

Hence, the time dimension is still the most reliable contextual information and some state-of-the-art solutions relied on the hypothesis that changes in the social context can be captured from the time as well, because users are likely to follow routine time patterns [45]. This assumption could potentially be applied to moods, since in some cases it may be affected by time, as identified in the pre-study. Therefore, a recommender system should utilise at least the time dimension to account for different contexts of use (TCX).

5.1.3 User Feedback

As mentioned in Section 2.3, user feedback can be divided into *explicit* and *implicit* feedback. Recommender systems more commonly rely on explicit feedback, e.g. in [28, 41, 53, 85], because it is generally more reliable and accurate since it comes directly from the user, while implicit feedback is system's inference about user's motivations and preferences based on indirect observations of their behaviour [34, 61, 79]. Thus, implicit feedback shows frequency of actions and indicates the *confidence* in the user-item relation-

ship, while explicit feedback shows the *preference* in it [34].

Therefore, there is a trade-off in choosing between these two types of feedback. In real-world cases most feedback is implicit, allowing more data to be collected about it, but it is noisy [34, 66]. On the other hand, explicit feedback has little noise, but is usually very sparse due to user's disinclination to rate items or the interaction cost to give feedback is too high in comparison to its utility [29, 34, 61], especially with a remote control.

Additionally, implicit feedback is especially suitable for making recommendations for series, since interaction with them is a recurring event [34]. However, in contrast to explicit feedback, it suffers from a high asymmetry between positive and negative feedback, since essentially interactions with an item are assumed to be positive [34, 66].

Nonetheless, implicit feedback has been found to improve the accuracy of predictions [2]. Therefore, it is beneficial to analyse user's behavioural data and model their implicit feedback. From the user experience perspective, Knijnenburg et al. (2012) [40] made a surprising observation that users had higher privacy concerns when being told that their explicit ratings are used to make recommendations than when given information that all behavioural data is used, which is in its essence more intrusive. The researchers discuss potential reasons, including the possibility that the effect of rating items on user's privacy is clearer, or that the users may simply forget that their behaviour is continuously analysed.

Therefore, implicit feedback seems to be more suitable for training a recommender system in terms of the user experience and how much data it provides, however more indicators of negative feedback should be collected to improve its reliability (IFB).

5.2 Analysing Existing Data

This section summarises findings based on analysing and exploring data available in Viihde, which was used as an example of data we could obtain from an operator. The goal of this investigation was to examine what problems could potentially arise when the service is offered to different operators and when dealing with data from a real-world VoD service. Moreover, I had limited access to data and some general information was obtained through interviews from Viihde's developers, in which cases I did not examine the underlying code or data myself.

5.2.1 Metadata

Following FIX to use feature-item information for explanations and, consequently, the recommender system, it was important to analyse what metadata a potential operator could have.

The video metadata I received contained various fields, such as the title, plot description, genre, country of origin, release year, duration, cast, directors, IMDb ratings, awards, keywords and mood. Initially, keywords (describing main themes or topics in a video) and moods field seemed to be suitable for the Category carousel together with the genre, however, the analysis uncovered that keywords were missing from almost 30% of movies and series, and moods from almost 60%. Moreover, these two fields were often very granular, hence in some cases assigned only to few titles, e.g. "lentokoneen maahanputoaminen" (en. "airplane hijacking") or "nörttien voitto" (en. "victory of nerds). Hence, in order to have categories other than plain genres, there would have to be a more universal way to extract categories that could work for different operators.

Furthermore, data inconsistency was a problem for other metadata fields. For instance, the main cast and directors were missing for over 15% of movies and series, and IMDb ratings for almost 30%. One of the main causes of this problem was that the video collection contained a lot of domestic (Finnish) content, many of which did not yet exist in IMDb's databases. This is a crucial problem in the context of the project, since operators may come from different countries or offer niche or small production content, which may not have all fields available. While a certain degree of metadata availability must be ensured to support MDD and MDB, their quality may not be guaranteed. On the other hand, basic metadata information, such as genre, country, plot and year, were found for most titles. Therefore, the inclusion of those features in the recommender system may be necessary for the context of this project.

5.2.2 User Data

Currently, Viihde does not have a rating system, therefore there is no explicit feedback from the users. Implicit feedback is based on what videos a user has watched in the past. However, relying only on those may not be sufficient for modelling implicit feedback (IFB). While opening a video may indicate that some characteristics of it captured the user's interest, it is difficult to identify whether the user actually liked it or what did they liked about it.

In terms of playing a video, an indicator of where the user stopped watching a video is stored for the purposes of Continue watching feature. Based on that it may be difficult to tell *how* a video was watched. For instance, a user

may have watched a video until the end and then rewind it, in which case the data would only indicate that the user stopped watching during the video, which would be in this case false. Therefore, the data about user behaviour lacks indicators of negative feedback and utilising video player data in the future could be useful for that purpose (IFB).

5.3 Implementation Suggestions

This section explores more concrete possibilities and examples of how a recommendation system for a VoD system could work, based on findings from the pre-study and the literature review discussed so far in this thesis. The suggestions are based on or inspired by solutions found in existing research. Firstly, in Section 5.3.1, I suggest how to extract different types of categories based on basic metadata information, such as genres, that should be available for most videos in a VoD system. Next, example approaches from the literature to diversifying content are presented in Section 5.3.2. Section 5.3.3 discusses the importance of having user profiles and how they can be initially configured. Suggestions for what metadata, user's behavioural data and contextual information should be collected can be found in Section 5.3.4. Lastly, suitable algorithmic approaches are discussed in Section 5.3.5.

5.3.1 Selecting Content for Category Carousel

In Chapter 4, I proposed using a Category carousel for browsing dynamic categories. The challenge in creating a Category carousel is in being able to offer categories that represent features of videos beyond basic metadata information, in order to satisfy the complex nature of preferences and diversification of content (CTP). For instance, some participants in the pre-study expressed to prefer videos with a "light" storyline in certain moods (see Section 3.7.2). However, the depth of a storyline is not typically in metadata information. Text-mining plot descriptions to extract qualities about a video could be a potential solution in the future, however, currently it would be a challenge to support reliably languages used by operators from different countries.

Eggnik and Bland (2012) [18] discovered that people correlate adjectives happy, humorous and light-hearted with comedy and sit-com genres. Hence, displaying videos belonging to those genres in a "Light storyline" category could serve as a starting point. Following the idea of extracting categories based on genres (which were identified to be a metadata field available for most videos in Section 5.2.1), I looked into the definitions of 10 main video

genres from multiple dictionaries and websites [3,4,5,6,7]. Figure 5.3 illustrates genres, keywords and phrases commonly used to define them, and how they are related to each other.

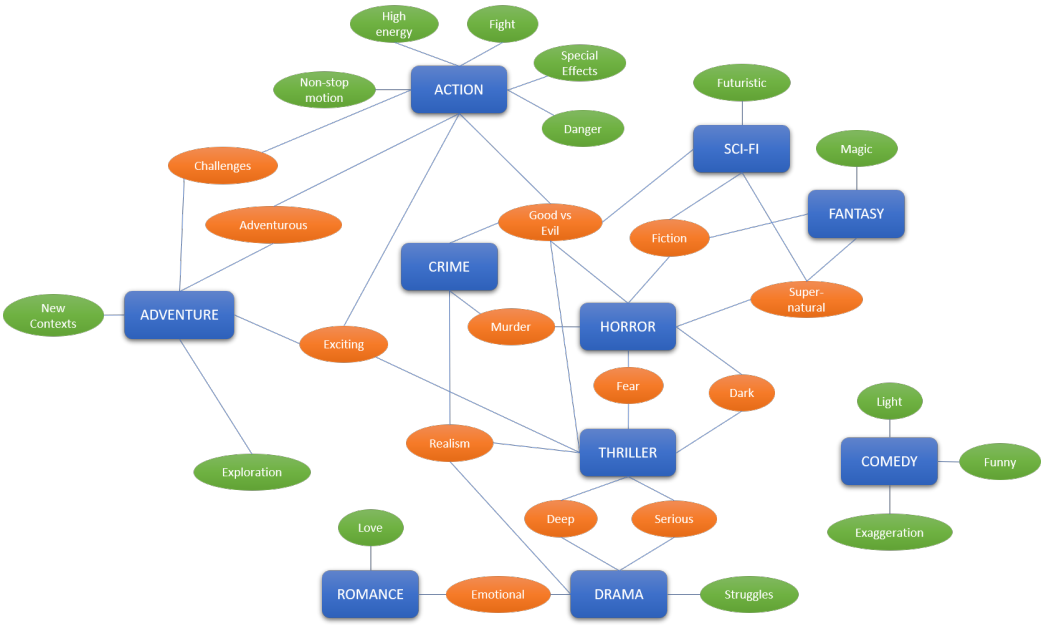


Figure 5.3: Main video genres and their relationship based on keywords most often used to define them. Orange nodes indicate themes that are common for two or more genres.

This "map" of genres and connections between them could be further refined, however, this is not in the scope of this project. Nonetheless, it demonstrates that many genres have common characteristics which could be used to create rules for new categories. For instance, videos from the action, crime, horror, sci-fi and thriller genres could belong to a "Good vs. Evil" category, or "Deep storyline" could be composed of dramas and thrillers excluding those that belong to comedies, which are distinct from other genres.

Therefore the category carousel could be composed of video features indicating:

- Depth of the storyline (e.g. initially extracted from genre metadata),

[3]https://www.filmsite.org/genres.html
[4]https://dictionary.cambridge.org
[5]https://en.wikipedia.org/wiki/List_of_genres#Film_and_television_formats_and_genres
[6]https://www.encyclopedia.com
[7]https://pediaa.com/

since it was found to play a vital role when it comes to mood. Examples: "Light storyline" and "Deep storyline".

- Quality, since it was an important criterion for some users. Additionally, higher quality was preferred in different contexts of use by some participants in the pre-study. The quality could be extracted from award or rating information, either international or local. Examples: "Oscar winners", "Top rated on IMDb", or "Critically-acclaimed".

- Type of plot or characters, e.g. initially extracted from genre metadata as shown in Figure 5.3, to describe different traits of a video. Examples: "Supernatural" or "Funny".

- Other available metadata information that is familiar for many people. Examples: "Psychological thrillers" (genre-based), "Domestic" (country-based), or "Recent releases" (year-based).

Moreover, the categories could represent one feature dimension, or multiple of them, e.g. "Top rated in 2019" indicating both the quality and the year of release. In the future, the recommender system could generate new categories based on the user's watching history.

To better support category-based browsing, it is essential that the recommender system works on at least two levels, similarly to Netflix (see Section 2.4.1). Firstly, the system must select categories suitable for the user ("category recommendation"). Secondly, since some categories may represent subjective terms (e.g. "Light-hearted", "Scary"), what they mean to each person may differ. Therefore, the recommender system should personalise what contents are displayed in the Filtered carousel for each selected category ("within-category recommendation"). Having both the categories and videos tailored to the user can improve content discovery significantly, as the user can focus on categories and filtered videos more relevant to them (CLA, CTF).

5.3.2 Diversifying Content

As mentioned in Section 5.1.1, empirical research showed that users may prefer recommendations from systems with worse prediction accuracy if they were diverse [12, 40]. Moreover, providing content diversification is important to support changing preferences in different contexts of use (CTP). Therefore, the system should strive to keep a balance between suggesting similar (exploitation) and novel (exploration) content. One popular approach to

achieve this, introduced by Ziegler et al. (2005) [90] is called *topic diversification*. It is applied on top of the recommendation list generated by the system and uses an *intra-list similarity* metric to evaluate the diversification level of the items in the list and select only those that have low similarity based on their features, e.g. genre. The researchers found that, even though topic diversification lowered the accuracy of the original models, the user satisfaction was notably higher.

A simple approach used in YouTube [17] is to constrain the number of recommended videos that are associated with the same feature. The authors also note that topic clustering or content analysis can be used for more sophisticated diversification of content.

It is important to note, that regardless of the chosen diversification technique, the system should apply it for both category and within-category recommendations to ensure that the user can explore new categories and videos when their preference changes (CTP).

5.3.3 Introducing User Profiles

TV is a shared device and even users within the same household may have significantly different preferences (see Section 3.7.2). Therefore, a recommender system may be receiving a lot of noisy data if multiple people are using the same account, significantly decreasing the performance of the system and consequently, making content discovery more difficult.

Since automatically detecting who uses the TV is intrusive, e.g. detecting faces from a camera or voices (see Section 5.1.2), a less invasive solution is to allow users of the same account who have different watching preferences to create their own profiles. Moreover, at the very minimum, there should be a possibility to distinguish between an adult's and child's profile, since their watching habits and preferences are significantly different, e.g. young children tend to re-watch the same content repeatedly, which is not a common behaviour for adults [4]. Additionally, this would reduce the risk of exposing a child to content inappropriate for their age.

The recommender system can be furthermore improved by allowing profile configuration, e.g. when the user first creates it. Figure 5.4 demonstrates how different categories could be used for that purpose. Selected categories then could be displayed in the Category carousel when the user starts to use the service and be given higher importance in category recommendation.

Additionally, the initial configuration should not overwhelm the new user with a high amount of choices to avoid the overchoice problem from the start. This implies that the categories must be diverse and indicate different types of features in a video. The categories selected for the profile configuration

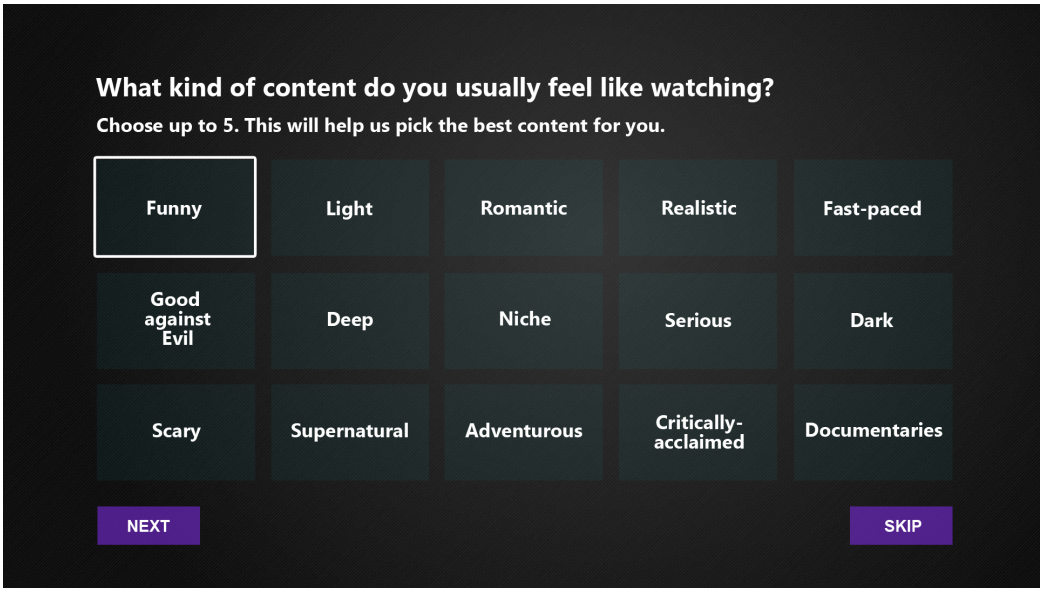


Figure 5.4: Category-based profile configuration screen.

screen shown in Figure 5.4 were chosen so that most important characteristics of videos are available and all main genres can be represented at least by one of the options. Moreover, the screen explains why this step is beneficial for the user to encourage going through it. This method of profile configuration was evaluated with potential users, and the evaluation results are described in the next chapter.

5.3.4 Choosing Data

Metadata

There is currently no standard for what metadata information should be used to model user's preferences for videos, however, using metadata in a recommender system crucial for its explainability (FIX). Moreover, it was shown that the addition of relevant metadata in a CBF algorithm can outperform a CF, which typically generates more accurate predictions [74]. Soares (2015) [74] compared the effect of using different metadata and their combinations on the quality and accuracy of CBF and CF algorithms, and discovered that the list of directors is the most discriminant factor. This can be explained by the fact that directors represent a certain style and quality, which a person may enjoy and not even be aware of.

However, as identified in Section 5.2.1, director metadata may not be always available, especially for different operators. Hence, to ensure a better quality of recommendations with the cost of increasing the computational complexity of the algorithm, the inclusion of more metadata is needed. Combination of genre, actors and directors metadata was found to provide high accuracy predictions [75]. Moreover, as identified in Section 3.7.1, year of production, country of origin and rating was important for some users.

Therefore, utilising the following metadata information for category recommendations is suggested: directors, genres, actors, year, rating and country. In terms of within-category recommendations, it is redundant to use metadata information indicating the category, since the within-category algorithm already considers only the subset of content filtered by that information. Moreover, it is not necessary to use rating and country metadata for within-category recommendations if director or actor information already define the category, as they are typically already correlated, e.g. the director represents a certain quality of video and actors tend to work on productions from the same country.

Behavioural data

Gathering data about user's behaviour in a system is crucial in modelling implicit feedback. Therefore, the following are suggested to collect:

(a) Consumption (watching) history, as it is the basis for any recommender system.

(b) Actions within the video player, especially fast-forward, and the length of parts that were skipped, since in most cases they can indicate that the user did not fully enjoy a video and had to skip the not interesting part.

(c) How long before the end of a video did the user stop watching. If a video was stopped long before the end and not continued within the next few sessions, it is an indicator that the user disliked a video.

(d) From which category a video was chosen. This would help to fine-tune the category recommendation. Moreover, in case there are duplicate items in different categories (which overall are to be decreased (CTD)) and the user chooses one of them, it is important to understand *why* was a video selected (for which features). For instance, if a video belongs to both "Oscar winners" and "Domestic" categories and was found from

the first, it shows that the user was interested in its critically-acclaimed feature rather than country of origin.

(e) How many videos were browsed within a category. Even if the user does not choose anything from a particular category, but browses through it extensively, it indicates that the user is interested in the features represented by the category, but the videos inside it may have not been too suitable and the within-category recommendation list needs to be modified.

(f) Whether a video is desired to be hidden from Continue watching. Giving the user an option to hide a video from Continue watching allows the system to collect essentially explicit negative feedback without the rating system. The users are also more likely to use this than a dislike option, as it has a clear and immediate effect on their item list.

Lastly, one of the main problems with implicit feedback is that there was an imbalance between positive and negative feedback (IFB). Points (b), (c) and (f) are used to create indicators of negative feedback and hence decrease that imbalance.

Contextual information

As discussed in Section 5.1.2, automatically inferring the current social and mood context non-invasively is still a research challenge. The time dimension becomes pivotal for context-aware systems (TCX) and needs to be therefore included to support changing watching preferences in different contexts of use (CTP). To decrease the complexity of the algorithm already affected by the need to utilise more metadata dimensions, the time dimension could be discretised to two variables, for instance day $D = \{workday, weekend\}$ and time of the day $T = \{morning, afternoon, evening, night\}$.

5.3.5 Algorithmic Approach

User experience and newly introduced regulations constrain what algorithms are suitable for a recommender system, as they need to be human-understandable and transparent (INM), which makes black-box algorithms unfit. Memory-based CF can use interpretable algorithms, however explaining why an item was recommended based on another previously consumed item (item-based CF) or other similar users (user-based CF) may be difficult to understand without a clearer cause of a recommendation, such as item's features.

Hence, rule-based and CBF approaches are the most understandable for an average user and allow using feature-item explanations (FIX).

A possible algorithm could be a hybrid of rule-based and CBF approaches. Since within-category recommendations are used to find important features and their combinations, rather than items themselves, a rule-based algorithm could be utilised and an association mining module would extract rules for generating categories. On the other hand, a CBF algorithm could be used to generate the list of items for within-category recommendations. For that purpose, cosine similarity is suitable for features where the order of its sub-features does not matter (e.g. "Action and Adventure" is equivalent to "Adventure and Action"), and Inverse Rank Measure otherwise (e.g. for list of actors or directors where their order may be important) [74].

Moreover, to support CTD, the recommender system needs to remove duplicated items (videos that are suggested in multiple categories and those that were already watched recently) from the within-category recommendations. A simple solution to remove duplicates of a video which was not yet watched is to first assign an importance measure to categories (e.g. how often the user browses it) and suggest that video only in the category with the highest importance to increase the chance of it being discovered. In case of videos that were already watched recently, they should be completely removed from the possible recommendations.

In terms of the frequency of updating the categories and videos (CTC), it could be adapted to how active a user is. For instance, a heavy user could receive changes twice a day, while an occasional user a few times a week. What is important in this aspect, is that the changes should not be drastic or large to reduce the confusion and the feeling of being completely lost in the system.

Chapter 6

Evaluation

6.1 UI Design Evaluation

One of the four fundamental phases in the user-centred approach to designing interactive systems is to evaluate the designs [36]. Hence, before moving to the implementation phase, it was important to investigate UI designs proposed in this thesis. The focus of this evaluation was on the wireframes presented in Chapter 4 and the profile configuration for the recommender system discussed in Section 4.2.2, and how well do they support content discovery from the perspective of the user interface (RQ1). Therefore, the main research questions here were:

- EV-RQ1: Do users understand how the proposed system works?
- EV-RQ2: How much perceived effort does it take to discover content with the proposed designs?
- EV-RQ3: Can users match proposed categories with their watching preferences?

6.1.1 Method

When evaluating designs, an essential consideration is the fidelity of a prototype. *Low-fidelity* prototypes are simple, cheap and fast to create. They are useful in demonstrating the main concepts and identifying issues related to the content and structure of a service, but not suitable for exposing errors or flaws in navigation, as they are not functioning products. On the other hand, *high-fidelity* prototypes are functional and interactive, giving the look and feel of the final product, however, are more expensive to develop and inefficient for proof-of-concept (PoC) designs. [62, Chapter 11]

Thus making a low-fidelity prototype was the most suitable for our project for multiple reasons: we were in an early stage of the development and also constrained by time, the research questions we wanted to answer with the study were on a conceptual level, and most importantly, a high-fidelity prototype is essentially the end result of the PoC project. To manifest the designs we decided to create *paper prototypes*, since we already had the designs created digitally and it was fast and cheap to print them.

To make the investigation of EV-RQ1 and EV-RQ2 more valuable, the users interacted with the paper prototype by saying out loud commands, limited to those available on a remote control with a D-pad, to navigate and take actions. Different states of the Home screen and the profile configuration were printed on separate pages. To show the user the intended look and feel of the system, we used refined designs made by the visual designer based on the wireframes shown in Figures 4.1 and 4.2. Figure 6.1 shows printed and refined versions of the wireframes.

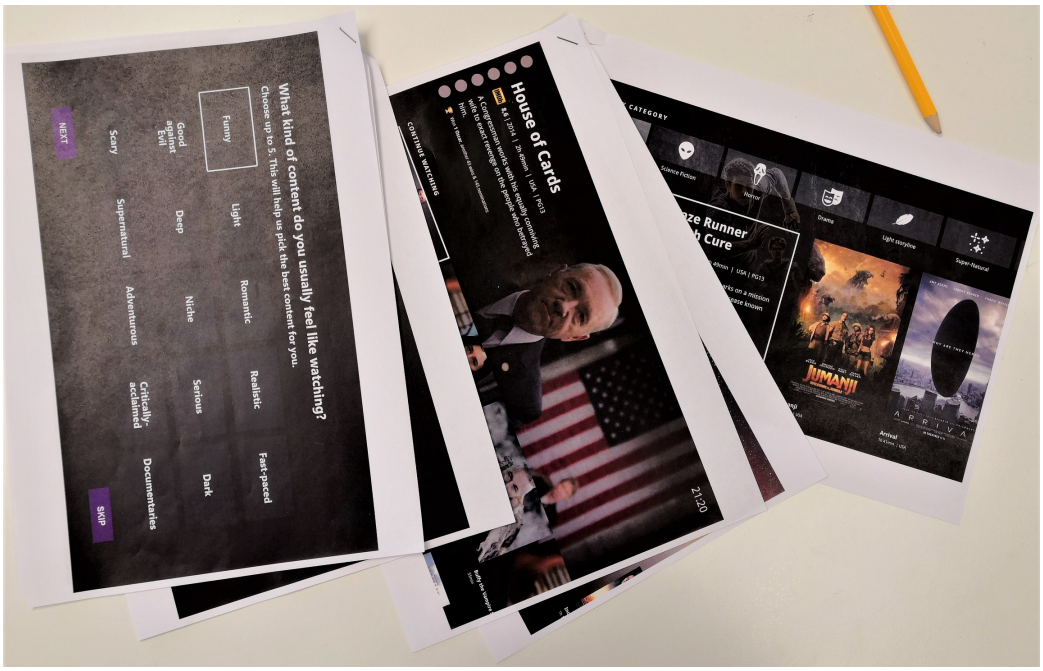


Figure 6.1: Printed paper prototypes of the proposed interface design used for evaluation. From left: profile configuration, Home and Category carousel screens in different states.

To triangulate, participants were asked to think-aloud as they interact with the paper prototype, and additionally, short semi-structured interviews

were conducted to ask explicitly about EV-RQ1 - EV-RQ3 and their overall feelings about the system.

6.1.2 Study Design

In this user study, each participant evaluated three screens: the profile configuration, Home screen and one additional screen, which was not in the scope of the thesis, hence its evaluation results are not discussed in this work. The screens are presented in the results, Section 6.1.5.

Tasks

To further investigate the feasibility of the category-based profile configuration, I decided to compare it with a thumbnail-based approach that Netflix uses. Moreover, to support MDA, I created an alternative profile configuration screen which also shows a description block under each thumbnail. Therefore the profile configuration screen was tested in three conditions: categories, thumbnails only and thumbnails with descriptions.

In terms of the main screen and category carousel, participants were firstly asked to briefly explore and navigate through the screens to observe EV-RQ1. Next, to investigate EV-RQ2, participants were given a task to find some funny content to watch.

6.1.3 Participants

Due to limited time resources, the most suitable sampling approach was *convenience sampling*. Potential candidates were approached within the company's premises, in the cafeteria areas and kitchens, to not disturb others' work. They were asked if they would like to look into and evaluate a new concept of a VoD service. 5 people (3 females) aged between 24 to 56 volunteered in the study. 2 participants had never used VoD services, the other 2 used them on a daily basis and 1 occasionally, a few times a month. There was no monetary incentive.

6.1.4 Procedure

The experiments were carried out in Elisa's headquarters in Helsinki, Finland, directly in the areas were participants were approached. The study was conducted during late afternoon hours when the company was less occupied and therefore the risk of disturbance was lower. The experiments took on

average 20 to 25 minutes (including testing of the screen which was not part of the thesis).

At the beginning, participants were briefly introduced to what the service under investigation is and the types of tasks in the experiments. Then, participants were asked to provide information about their age and familiarity with VoD services. All participants agreed for the sound to be recorded throughout the experiment for later analysis.

Firstly different configuration screens were tested in a random order. After "configuring" their profiles with all three types, participants were asked to pick their favourite and least liked configuration method, explain their choices and describe how easy or difficult it was to use them.

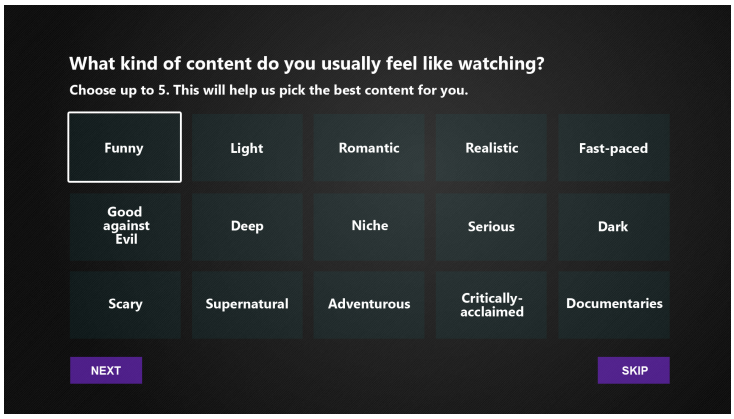
To interact with the Home screen and the Category carousel, participants were asked to say what actions they want to perform using four directions or "OK" to confirm, and how do they expect the system to respond. I was moving a pointer on the paper screen according to their verbal commands if they were valid and informed that the action was not possible otherwise. After testing the Home and category screen, participants were asked to explain how they felt about discovering content with the proposed system, what did they like about it and what not.

After all tasks, there was a short closing interview, where participants were asked for overall feedback, if they would change anything they in the proposed designs and whether they would use the actual product in the future.

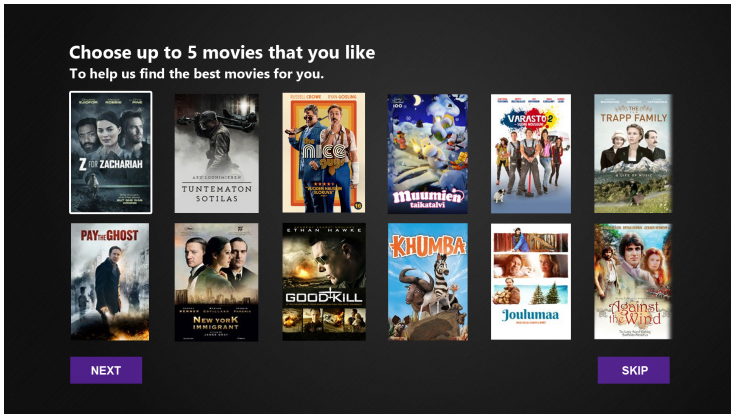
6.1.5 Results

Profile Configuration

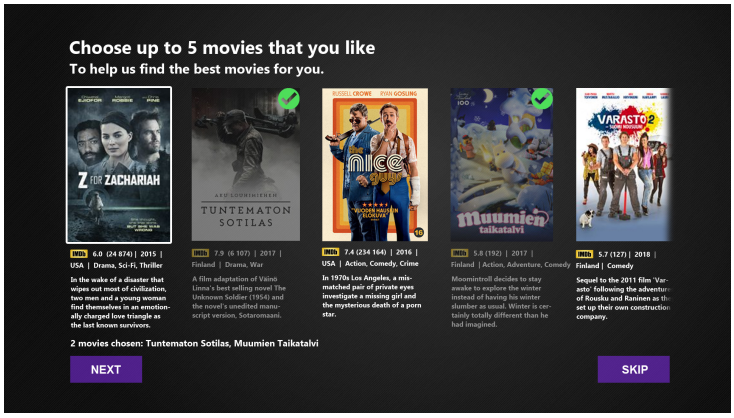
Figure 6.2 shows designs of three types of profile configuration that were tested in the study. 4 out of 5 participants preferred to use category as a base for setting up their profile than the other two methods. They reported that this method was much easier and they could quickly find categories that they were interested in. The remaining participant did not understand multiple words from category names due to language difficulties and reported it as a reason to prefer the thumbnail-based method. For instance, the terms "Deep" and "Dark" were too vague for that participant and more descriptive names, like "Deep Storyline" or "Dark Tone", would have been preferred. In terms of other category names, 2 participants did not understand what "Critically-acclaimed" meant in English and 1 asked what was "Niche". Therefore, the problems that occurred with the category-based configuration did not lie in the method itself, but in the language.



(a) Category-based



(b) Thumbnail-based



(c) Description-based

Figure 6.2: Interface designs of three types of profile configuration.

The thumbnail-based configuration type was rated as the least liked by 4 participants. The main reason was that they were not familiar with presented titles and there was not enough information provided about them to make a choice. Using the description-based configuration was easier for them than the one with only thumbnails. This confirms the need for MDA in any view where the user can browse through different videos. The participant which preferred the thumbnail-only configuration was already familiar with most titles that were displayed. However, that participant also reported that they chose only the titles they knew, but not necessarily the ones they liked. This shows that the lack of additional information may lead to *buyer's remorse*.

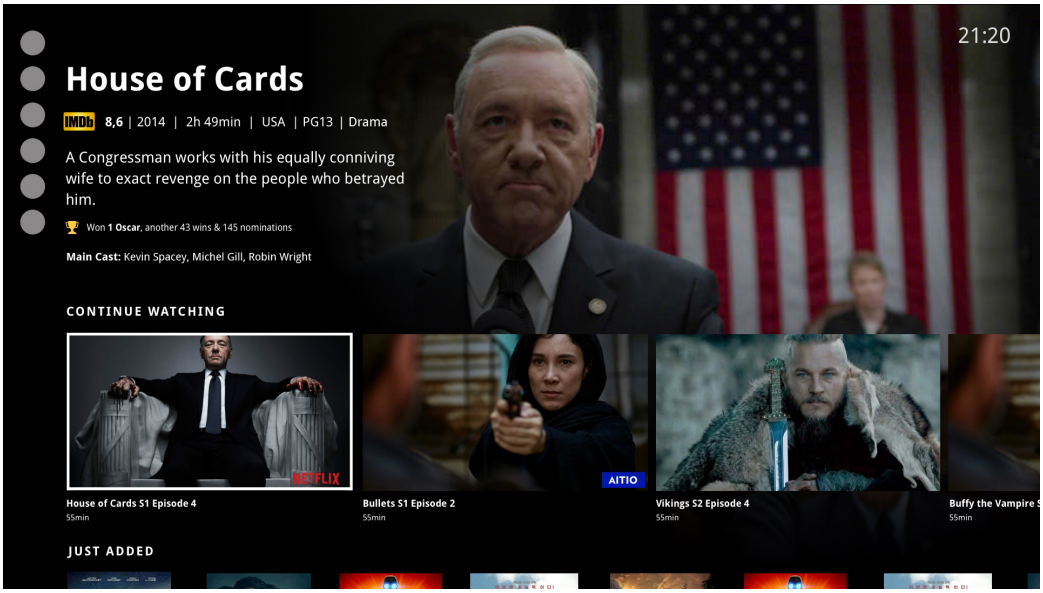
However, the main problem with using videos as means to set up a profile (thumbnail-based or description-based configurations) was that participants felt that it was too specific and would not describe the broader range of their interests. Therefore, the category-based profile configuration was found to be the most suitable in indicating new users' general preferences. Most participants were also able to spot at least 3 categories that they would be interested in watching almost immediately, which suggests that browsing by category with the Category carousel has a high potential to make the content discovery faster. Moreover, choices made in the category-based configuration could be used to train directly the category recommendation algorithm, allowing the service to be better personalised already from the beginning of use.

Home screen and Category carousel

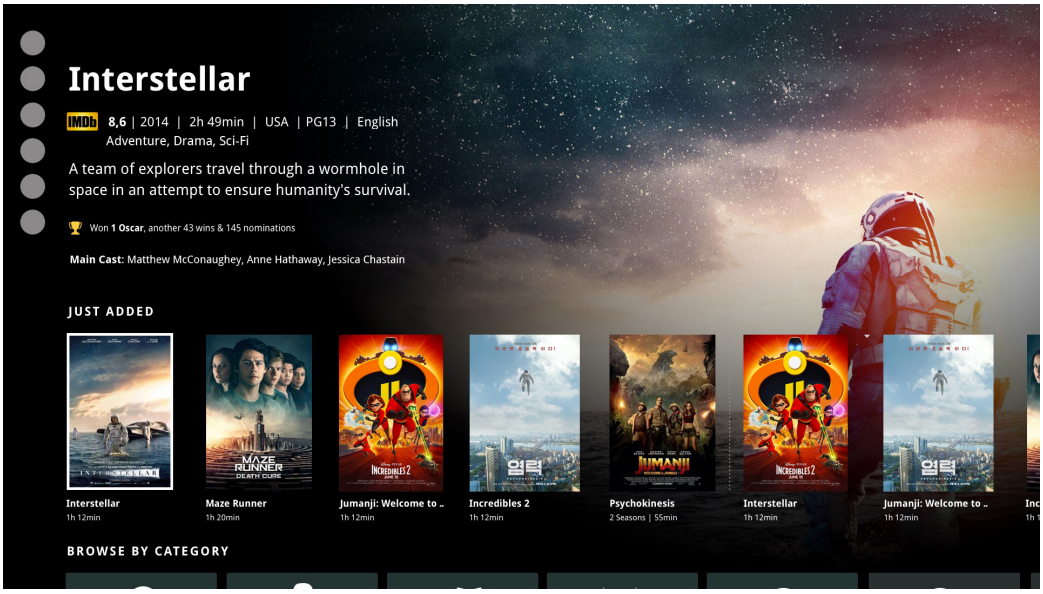
Figure 6.3 presents the designs of the Home screen used in the study. None of the participants had problems in the navigation or identifying the current focus point (FPD). The only issue that was observed during the study was when 2 participants initially tried to find a way to go back to the profile configuration screen with categories from the Home screen. However, this behaviour clearly indicated that they preferred to browse based on categories rather than carousels.

All participants liked the Category carousel, shown in Figure 6.4, and diversity of the available categories. They reported that the Category carousel was easy to use and understand. The following are comments from different participants regarding the Category carousel:

- *"I didn't imagine there would be a "Browse by category". It looks good. (...) Modern when I compare that to Viihde."*
- *"For me it's kind of natural to browse by category."*



(a) First static carousel



(b) Last static carousel

Figure 6.3: Interface designs of the Home screen made by the visual designer based on wireframes shown in Figure 4.1.

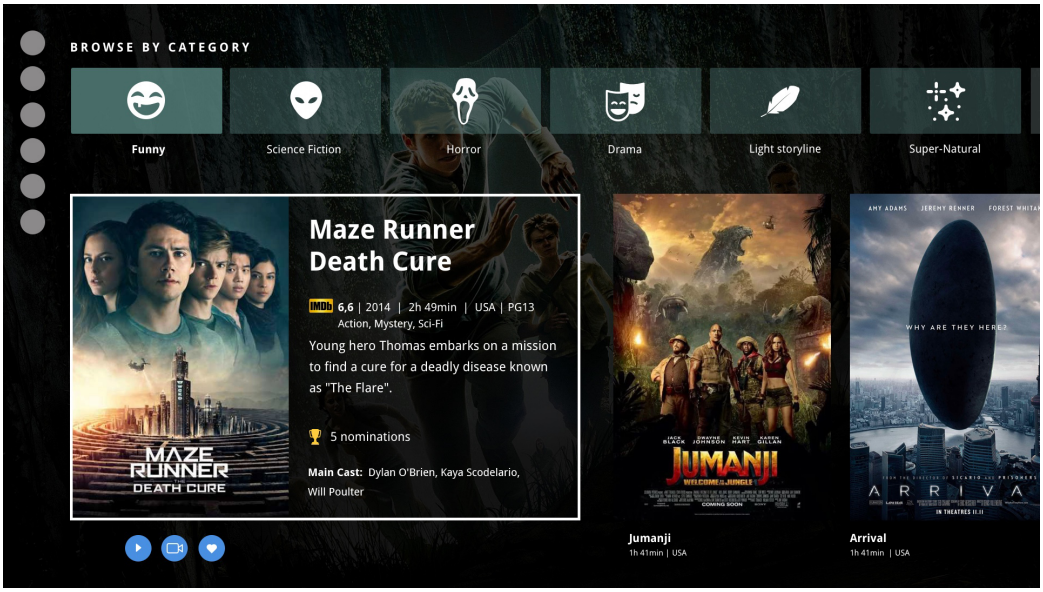


Figure 6.4: Interface design of the screen with the Category carousel made by the visual designer based on the wireframe shown in Figure 4.2.

- *"I think it's good (...) shows what kind of movies and categories you have there, it's more clear. This way the service would be easier even for kids to use."*
- *"For some people, like me, the category browsing is very useful for day-to-day use."*
- *"I don't see any ways to makes this easier to use."*

This shows that CLS, CTP and CTF were achieved with the proposed design.

With regards to descriptive information, 3 participants commented that the proposed system shows enough information for them to make a decision. 1 participant reported to have wished to see more information about the characters and plot, and 1 participant wanted to read more about particular seasons of a series. Hence, MDB and was met and also MDD for most participants. The video view containing more information about it and the series view were not in the scope of this thesis, but the feedback was valuable for its development in the future.

Overall Feedback

Results were very positive and show that the proposed UI design of the service was perceived to be simple and easy to use. Participants who had experience using Netflix, Viihde and other VoD services said that with the proposed design there is less browsing and effort needed to find content when compared to the services familiar to them, and some praised the different approach to content browsing. 3 participants were interested to know more about the service and expressed the desire to use it when it is released. The remaining two participants liked the UI designs, but would decide to use the service depending on what content would be offered in the actual system.

6.2 Company Feedback

Findings from this work were presented in front of multiple teams, receiving positive feedback. Furthermore, the guidelines for user interface and recommender system design were evaluated by the Viihde team (the guidelines are summarised later in Section 7.1). The evaluation revealed that most of them resonate with the current work and plans, validating the overall usefulness of the guidelines. The remaining guidelines started discussions on their potential applications in Viihde. The Viihde team commented that the guidelines provide important insights and are useful in aiding the design decisions and could help in aligning the business, development and design goals.

The team identified that with the current implementation of the Viihde on the set-top-boxes, there is no trivial way to track user's interaction with the carousels. Therefore, identifying from which category a video was found would be challenging. This, however, is specific for that case and does not necessarily apply to implementations of other systems. Other than that, developers were positive about the possibility to use video player's data, such as fast-forward event, for modelling the implicit feedback. The team also made a remark that, in a multi-platform service such as Viihde, device information would be also helpful in identifying different contexts of use.

Lastly, the team I was part of commented that conducted studies and involving users in the design process were very helpful and valuable in aligning the design decisions to user needs and improving the user experience. Furthermore, the team found methods selected for the user studies appropriate and effective in identifying pain points in existing services and the needs of users. The proposed UI designs were accepted for implementation in the PoC project and the outcome product was well received within the company and by its evaluators, and it was approved for further development.

Chapter 7

Discussion

The primary purpose of this study was to define guidelines which would help in designing a video-on-demand service for a TV interface that allows discovering content with perceived less effort. The main motivation for this study was that nowadays users are presented with large amounts of options in those services, which on one hand make the service appear more attractive to the user, as it has more to offer, but on the other hand can overwhelm them. This project was carried out as part of a project to build a proof-of-concept VoD service in Elisa.

Earlier research suggests that when presented with a lot of choices, the user may suffer from the problem of overchoice, which is likely to result in paralysis analysis or buyer's remorse [28, 56, 72, 78]. The main solution for this problem is to introduce recommender systems into a service to display a subset of content to the user that they are likely to appreciate, based on their past preferences, e.g. in [16, 17, 25, 41, 90]. However, in the context of content discovery, the main focus in the research has been on improving the prediction accuracy of different algorithms, but less attention has been put on the user experience of those systems [40, 64]. Hence, this thesis took on a user-centric approach to improving the content discovery process in VoD systems, which is especially crucial for entertaining services. This work investigated the aforementioned issue not only from the algorithmic perspective of the underlying system, but also the user interface and interaction.

The main findings of this work reveal that the most popular way of displaying video content on TV - within carousels placed sequentially on a page - is affecting negatively the content discovery process for multiple reasons. First of all, the interaction cost is high due to an excessive amount of page scrolling required to browse through content. Consequently, scanning through available categories of content is very inefficient, as in most popular services, like Netflix or Viihde in Finland, the user typically can see at most

two categories at the time. This is very crucial, since providing many diverse categories is needed due to the complex nature of preferences for watching TV, which can change within a single person depending on the context, e.g. who they watch TV with or what mood they are in. Furthermore, metadata information about an item is essential for the user to make a decision on what to watch, however it is not always provided or is not sufficient enough. Moreover, while the underlying recommender systems may have high predictive power, their usefulness is decreased if the users cannot find meaningful explanations as to why a service would imply that the recommended items are relevant to them. Therefore, it is clear that in order to provide an enjoyable content discovery experience, the design needs to be considered also in terms of how the content is displayed to the user and the cost of interaction with the interface it requires.

To answer RQ1 on *how to present information on a TV interface to support efficient content discovery*, a list of design implications was made based on a user study, which then guided the design of the new interface. Instead of having a long, sequential view of carousels, I proposed to utilise a 2-level carousel, called Category carousel, as the main method of browsing available categories and content in the service. Furthermore, the type of interaction Category carousel enforces is similar to filtering the content based on a category dimension, giving users the sense of being in control of what they see. Moreover, in the proposed design, it was ensured that key information about the currently selected video is always provided. Additionally, the information is chunked into differently styled groups for readability and to allow easier spotting of relevant parts. Lastly, since users often want to understand why something is recommended to them, it was important to provide brief explanations on the interface, by referring to the previously watched items the recommended one is related to and features that they share (feature-item explanation).

In terms of RQ2 regarding *what kind of recommender system approach could fulfil user's needs in the context of content discovery*, the necessity to be able to explain recommendations to the users was the main constraint on the possible algorithms. Using machine learning to explain black-box algorithms, which typically offer high prediction accuracy, but are not human-understandable, has become a new research field in the area of XAI [27]. However, while there have been promising results in using an algorithm to explain results of another one, this approach cannot guarantee transparency, which became a critical issue since the recent introduction of new regulations, e.g. the "right to explanation" in GDPR. Therefore, to satisfy both the user experience of recommender systems and legislations, it is advisable to use interpretable models, such as rule-based algorithms. Moreover, the

recommender system should strive to account for contexts [16, 45, 58] and diversify content [12, 40, 90] to support the changing and complex nature of preferences. Furthermore, to better support content discovery, the recommender system should work on two levels, similarly to the Category carousel: by providing both category and within-category recommendations. Lastly, it is beneficial to model user's implicit feedback, since users typically rate a very low percentage of the content they watch [29, 34, 61] and explicit feedback is a subject of privacy concerns [40]. The main problem with implicit feedback is that it is essentially providing only positive feedback (e.g. if a user watched a video, it is assumed that the user likes it). To counter that, I suggested collecting indicators of negative feedback, such as fast-forward actions and how long before the end was a video stopped watching.

The interface design of the proposed system was evaluated with users and received positive comments. Participants especially liked browsing with the Category carousel, as it was simple and gave them the impression that they do not need to browse much. Therefore, this shows that the goal of creating an enjoyable experience of content discovery was achieved with the proposed design, which was guided by the design implications for RQ1. The design suggestions for recommender system, answering RQ2, were evaluated by multiple developers and found plausible, with the exception of the possibility to gather all the suggested implicit feedback due to limitations in the current implementation of the Viihde system, which does not necessarily apply to other VoD systems.

Lastly, the main findings were summarised into 12 design guidelines in the next section. They were evaluated by the Viihde team, which found them viable and useful in aligning the business, development and design goals. Even though the guidelines were not evaluated on a high-fidelity functioning product, the insights from this work were showed to be conceptually useful for designing an efficient content discovery experience and therefore can be used to guide its design and implementation from a user-centric perspective.

7.1 Design Guidelines

This section summarises the key findings of this work into 12 guidelines that aim to help in designing an efficient content discovery user experience, in relation to user interface and recommender system designs. The guidelines are grouped by the main research questions of the thesis.

RQ1. *How to present information on a TV interface to support efficient content discovery, given the limitations of the input device and the layout possibilities?*

1. Provide key information about the items consistently in all browsing views for easier comparison of items.
2. Group key information into meaningful chunks and create a consistent visual hierarchy to allow faster scanning and identification of relevant information (e.g. using Gestalt's laws of proximity and similarity).
3. Organise items into higher-level categories and minimise the number of page loads needed to scan through them to give the user a better understanding of the type of available items and means to quickly narrow down (filter) to the desired ones.
4. Ensure that the search functionality is easily accessible from any screen and its capabilities are clearly communicated.
5. Show an explanation of why an item is recommended and use simple feature-item references to increase recommendations' meaningfulness in the decision-making process.

RQ2. *What kind of recommender system approach could fulfil the needs users have in the context of content discovery?*

6. Use human-interpretable models to ensure high explainability and transparency of the underlying algorithm, important for better user satisfaction and required by some regulations.
7. Minimise the number of duplicated items and those recently consumed by the user from recommendations, to avoid artificially increasing the catalogue volume that the user browses.
8. Strive to differentiate between different contexts of use, e.g. by utilising the time dimension, device information or location.
9. Personalise content on two levels: category and within-category recommendations, to suggest the most relevant categories to the user and the items inside them.
10. Collect indicators of negative feedback when modelling implicit feedback to keep a balance between the indicators of positive and negative feedback.

The following are guidelines which relate both to RQ1 and RQ2:

11. Diversify key information, categories and items that are displayed to the user, to support changing preferences and needs in different contexts of use.

12. Introduce changes in the recommendations gradually to avoid disorienting users and better support their learning of the content.

It is important to note that, while this research focused on VoD services and TV interfaces, above-listed guidelines can be applied beyond those two cases to any services in which a user is presented with a large quantity of items to make choices from, e.g. in online shopping.

7.2 Limitations

The main limitation of this work is that both the implication gathering and evaluation studies were on a small scale, therefore may not be generalised. Moreover, most of the participants in both studies had higher education, which is not representative of the whole population of users of online services like VoD. However, some of the findings from the initial user study are on par with work from other researchers, e.g the need for explainable recommendations or diversifying content, increasing the reliability of results.

Moreover, designs could not be evaluated on the real system within the timeframe of the thesis to observe whether the resulting system can make content discovery more efficient and enjoyable for the users in practice. Furthermore, due to time limitations, it was decided to run laboratory studies, however, they cannot uncover the full context of use, such as family interactions or other external factors that may affect the content discovery process and should be accounted for in the service design.

Lastly, all analysis was performed by one person, while studies suggest that for the reliability of results, multiple investigators should participate in the analysis phase, e.g. to decrease the bias on the interpretation of results [62, Chapter 8].

7.3 Main Contributions and Future Work

In the past decade, researchers started to put more value into the user experience of recommender systems and emphasise its importance over the algorithmic prediction accuracy, e.g. in [12, 40, 64]. From the user interface perspective of content discovery, there has been research on designing

the search experience, for instance, in terms of the complexity and learnability of the query syntax, highlighting and labelling results, or combining search with manual filtering to narrow down the results [54, 68]. However, to my knowledge, there is no extensive research on the interface, interaction or information architecture design of content discovery when the user has no specific type or theme (e.g. genres) in mind, which is often the case when browsing content in video-on-demand services, as identified in the pre-study. Moreover, there are still no clear user-centred guidelines on how to design content discovery both in terms of the user interface and the recommender system, or their intersection. Furthermore, content discovery is significantly more difficult on a TV interface due to the limited interaction capabilities with the current technology. Therefore, in this thesis, I proposed visualisations of the interface design of a video-on-demand system suitable for a TV interface and discussed the design decisions in the context of content discovery from the perspective of information architecture. Furthermore, key learnings from the user studies and the review of academic research were collected into guidelines for designing content discovery from the perspective of the user interface and the recommender system. The guidelines were generalised making them applicable also in other types of online services, such as in e-commerce, and not only on TV interfaces.

In the future, the guidelines would need to be implemented into a functioning product and their feasibility evaluated on a larger scale with users, which was not possible within the limited timeframe of this project. If further research confirms the viability of the guidelines, they could potentially become requirements for user-centred design of content discovery.

Regarding the search functionality, which is an important part of content discovery, one of the guidelines suggests making it visible and easily accessible. Finding videos through search was a tedious task for many participants in the pre-study, especially due to typing, which is inconvenient and time consuming with a remote control. Moreover, using a remote control significantly limits navigational possibilities on TV interfaces. Nowadays, voice search and voice commands became widely available on many devices, e.g. smartphones, as they allow to search and use a system hands-free [11, 35, 55]. The main ongoing challenge in building voice interfaces is the unreliability of speech recognition in different languages and accents. For instance, in the recent study from Pyae and Joelsson (2018) [65] on the usability and user experience of Google Home conducted with 114 participants, issues with the recognition of non-English words and the lack of support for some of the participants' native languages were reported. Therefore, voice interaction may not yet be feasible for non-English services with the existing technology.

Another potential input method could be mid-air gestures, which have

been in the interest of many researchers ever since sci-fi movies introduced them. A known problem with this type of interaction is the *gorilla arm* effect, which refers to the fatigue and pain in the arms caused by holding them unsupported and for a prolonged time in the air [32, 80]. Since it requires more physical effort from the user, it is not clear whether mid-air gestures would be suitable for a leisure activity like watching TV, which in principle does not involve much movement from the user. Thus, alternative input methods for TV should be investigated further, especially in terms of their usefulness, effectiveness, suitability and accessibility (universal usability).

The last guideline regarding the gradual introduction of changes to the recommendations is also academically relevant for future research. The frequency and scale of changes in recommendations in Netflix were reported to be too abrupt and radical by some of its users in the conducted pre-study, which had caused them to "lose track" of the items that caught their attention during previous uses of the system, or specific categories they were interested in. Changes in the interface have been previously studied in terms of the layout and its elements, e.g. in the field of adaptive user interfaces [5, 9]. It is not known yet whether similar approaches would be suitable or optimal for updating recommendations. Though, it is certain that the frequency and scale of changes should be adapted to the user. For instance, some participants in the conducted user study indicated that they would like to see effects of their actions to be reflected in the system immediately (e.g. to never show again content that was disliked), and others were frustrated by the frequency of changes. Additionally, how the changes in the category and within-category recommendation would differ also need to be investigated in the future.

Explanations have been studied in terms of the usefulness of justifications they use and wording. However, an interesting question for future research is whether different phrasing of an explanation could intrigue and encourage users more to watch an item. As identified in the pre-study, people currently rely on and trust human recommendations more than those coming from an artificial system. Hence, it may be beneficial for the design of recommender systems to explore different human-like phrasing of explanations since they could potentially increase the trustworthiness in the quality of system recommendations. However, there may also be a risk that the human resemblance would elicit in the users eerie feelings, which is a phenomenon first discussed in robotics, called the *uncanny valley* [52]. Therefore the potential use and effect of human-like phrasing of recommendations is useful to research.

The largest knowledge gap identified in this research was the lack of solutions for automatic mood detection that would be suitable for TV watching, even though mood is one of the essential factors affecting the preference for

content to watch. The context-aware solutions proposed in the literature rely on the manual input of the context by the user [53, 73], which is not suitable for a leisure activity and even more difficult using a TV remote control. While there exists technology to detect people, e.g. through image recognition [51], and researchers have attempted to recognise emotions based on physiological measures captured from medical equipment, such as electroencephalography (EEG) monitoring devices [38], these solutions are either intrusive or inconvenient for in-home use. However, with the increasing popularity of health applications monitoring heartbeat or blood pressure in consumer devices, such as smartwatches, mood and emotion recognition for home entertainment activities may become more achievable as they are correlated with those measures, e.g. stress commonly increases blood pressure [76], and the devices are small and easily available for regular users. Hence, the feasibility of using physiological measures recorded from consumer devices for mood detection and the willingness of potential users to provide such data for improving recommender systems needs to be evaluated.

Lastly, as mentioned throughout the work, TV is a shared device and TV watching is often a social activity. This work has not tackled with the problem of multi-usage of a VoD system other than suggesting to allow creating profiles and diversifying content for different contexts of use. For instance, how does sharing a remote control, interaction with other people in the household, or other household activities, such as cooking or cleaning affect usage of the service and content discovery? Therefore, in the future, it would be valuable to conduct user studies in user's natural environment to discover the whole spectrum of factors influencing the content discovery process, or invite users with people they typically watch TV with to a laboratory study.

Chapter 8

Conclusions

This study was conducted as part of a proof of concept project to develop a video-on-demand service that allows its users to discover content with perceived less effort. In this study, I have identified the main pain points in discovering content in state-of-the-art VoD services by conducting a user study in a controlled environment, in which subjects were observed while using Netflix and Elisa Viihde, and also designed their ideal services. Findings from this study were translated into design implications for the user interface and recommender system of a VoD service, which emphasised the most important aspects that affect the content discovery process. Additionally, other usability problems identified in the studied services were also reported.

The design implications were further translated into wireframes to demonstrate how such a VoD service could potentially look like. The main features of the proposed interface are: descriptive information about the currently selected video is always provided, long pages with sequentially displayed video carousels were collapsed into a 2-level Category carousel, and information about recommended items always include a brief explanation why the system considers it to be relevant to the user.

In terms of recommender systems, the most important finding is that it needs to be human-interpretable, in order to provide transparency and explanations, which are desired by the users and recently also required by some regulations, e.g. GDPR. Moreover, explanations which best support decision making contain information about what other items the recommended one is similar to, and based on which feature. This further narrows down the possible algorithms mainly to content-based filtering and rule-based methods. Furthermore, since preference for content may change within a user depending on the context, it is necessary that VoD services provide many diverse types of content. Hence, the recommender system should determine both what categories should be shown to the user, and suggest content within

each category, while ensuring that the recommendation lists are diverse. To improve the performance of a recommender system, it is also important that behavioural data is collected to model implicit feedback.

The resulting interface design was evaluated by users. The Category carousel approach was well received by all participants and the service overall gave users the impression that they would not need to spend much time browsing content, validating that the goal of the work in terms of the user interface design was achieved. The main contribution of this work – the set of guidelines for user interface and recommender system design for improving the user experience of the content discovery process – was further evaluated by developers and designers at the company where the research was carried out, and no conceptual issues were identified. Moreover, the guidelines were thought as useful and insightful, and helpful in aligning business, design and development work. Therefore, the output of this work can aid a successful design of the content discovery process and moreover, they can be applied beyond the design of VoD services or TV interfaces. Lastly, the thesis discusses ideas and opportunities for future research.

Bibliography

[1] Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), ACM, p. 582.

[2] Abel, F., Bittencourt, I. I., Henze, N., Krause, D., and Vassileva, J. A rule-based recommender system for online discussion forums. In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems* (2008), Springer, pp. 12–21.

[3] Ali-Hasan, N., and Soto, B. 8 things to consider when designing interactive tv experiences. In *TVX 2015 - ACM International Conference on Interactive Experiences for Television and Online Video* (2015).

[4] Andersson, M. Enhancing content discovery in video on demand services for children, 2017.

[5] Basson, S. H., Kanevsky, D., and Oblinger, D. A. Predictive user modeling in user interface design, Oct. 20 2015. US Patent 9,165,280.

[6] Bernhaupt, R., and Pirker, M. M. User interface guidelines for the control of interactive television systems via smart phone applications. *Behaviour & information technology 33*, 8 (2014), 784–799.

[7] Bilgic, M., and Mooney, R. J. Explaining recommendations: Satisfaction vs. promotion. In *Beyond Personalization Workshop, IUI* (2005), vol. 5, p. 153.

[8] Biran, O., and Cotton, C. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)* (2017), vol. 8.

[9] Browne, D. *Adaptive user interfaces*. Elsevier, 2016.

[10] Cacheda, F., Carneiro, V., Fernández, D., and Formoso, V. Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Transactions on the Web (TWEB) 5*, 1 (2011), 2.

[11] Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., Wagner, D., and Zhou, W. Hidden voice commands. In *25th {USENIX} Security Symposium ({USENIX} Security 16)* (2016), pp. 513–530.

[12] Chang, N., Irvan, M., and Terano, T. Designing a hybrid recommendation system for tv content. In *Intelligent Decision Technology Support in Practice*. Springer, 2016, pp. 217–229.

[13] Choi, B., Lee, Y., and Park, S. S. A research of user interface design elements on smart tv. *ADADA - International Conference of Asia Digital Art and Design* (2014), 219–222.

[14] Chorianopoulos, K. User interface design principles for interactive television applications. *Intl. Journal of Human–Computer Interaction 24*, 6 (2008), 556–573.

[15] Collazos, C. A., Rusu, C., Arciniegas, J. L., and Roncagliolo, S. Designing and evaluating interactive television from a usability perspective. In *2009 Second International Conferences on Advances in Computer-Human Interactions* (2009), IEEE, pp. 381–385.

[16] Cremonesi, P., Modica, P., Pagano, R., Rabosio, E., and Tanca, L. Personalized and context-aware tv program recommendations based on implicit feedback. In *International Conference on Electronic Commerce and Web Technologies* (2015), Springer, pp. 57–68.

[17] Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems* (2010), ACM, pp. 293–296.

[18] Eggink, J., and Bland, D. A large scale experiment for mood-based classification of tv programmes. In *2012 IEEE International Conference on Multimedia and Expo* (2012), IEEE, pp. 140–145.

[19] Ericsson, K. A., and Simon, H. A. *Protocol analysis: Verbal reports as data.* The MIT Press, 1984.

[20] Foltz, P. W., and Dumais, S. T. Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM 35*, 12 (1992), 51–60.

[21] Frey, B. S., Benesch, C., and Stutzer, A. Does watching tv make us happy? *Journal of Economic psychology 28*, 3 (2007), 283–313.

[22] Fukukura, J., Ferguson, M. J., and Fujita, K. Psychological distance can improve decision making under information overload via gist memory. *Journal of Experimental Psychology: General 142*, 3 (2013), 658.

[23] Gibbs, G. R. *Analyzing qualitative data*, vol. 6. Sage, 2018.

[24] Gigerenzer, G., and Todd, P. M. *Simple heuristics that make us smart.* Oxfor University Press, New York, 1999.

[25] Gomez-Uribe, C. A., and Hunt, N. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS) 6*, 4 (2016), 13.

[26] Google. Android tv. https://designguidelines.withgoogle.com/android-tv/. [Online; accessed 7-April-2019].

[27] Gunning, D. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* (2017).

[28] Haddad, M. R., Baazaoui, H., Ziou, D., and Ghézala, H. B. Towards a new model for context-aware recommendation. In *2012 6th IEEE International Conference Intelligent Systems* (2012), IEEE, pp. 021–027.

[29] Harley, A. Individualized recommendations: Users' expectations & assumptions. *Nielsen Norman Group, Articles* (2018). Online at: https://www.nngroup.com/articles/recommendation-expectations/.

[30] Herlocker, J. L., Konstan, J. A., and Riedl, J. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (2000), ACM, pp. 241–250.

[31] Hertzum, M. Images of usability. *Intl. Journal of Human–Computer Interaction 26*, 6 (2010), 567–600.

[32] Hincapié-Ramos, J. D., Guo, X., Moghadasian, P., and Irani, P. Consumed endurance: a metric to quantify arm fatigue of mid-air interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), ACM, pp. 1063–1072.

[33] Hornbæk, K., et al. Some whys and hows of experiments in human–computer interaction. *Foundations and Trends® in Human–Computer Interaction 5*, 4 (2013), 299–373.

[34] Hu, Y., Koren, Y., and Volinsky, C. Collaborative filtering for implicit feedback datasets. In *ICDM* (2008), vol. 8, Citeseer, pp. 263–272.

[35] Hurst-Hiller, O., and Farago, J. Searching for content using voice search queries, Mar. 2 2010. US Patent 7,672,931.

[36] ISO. Ergonomics of human-system interaction – part 210: Human-centred design for interactive systems. Standard, International Organization for Standardization, March 2010. ISO 9241-210:2010.

[37] Kenthapadi, K., Le, B., and Venkataraman, G. Personalized job recommendation system at linkedin: Practical challenges and lessons learned. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (2017), ACM, pp. 346–347.

[38] Khalili, Z., and Moradi, M. H. Emotion recognition system using brain and peripheral signals: using correlation dimension to improve the results of eeg. In *2009 International Joint Conference on Neural Networks* (2009), IEEE, pp. 1571–1575.

[39] Kimani, S. *WIMP Interfaces*. Springer US, Boston, MA, 2009, pp. 3529–3533.

[40] Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., and Newell, C. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction 22*, 4-5 (2012), 441–504.

[41] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., and Riedl, J. Grouplens: applying collaborative filtering to usenet news. *Communications of the ACM 40*, 3 (1997), 77–87.

[42] Koren, Y. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (2008), ACM, pp. 426–434.

[43] Kulkarni, S. S., Reddy, N. P., and Hariharan, S. Facial expression (mood) recognition from facial images using committee neural networks. *Biomedical engineering online 8*, 1 (2009), 16.

[44] Linden, G., Smith, B., and York, J. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 1 (2003), 76–80.

[45] Liu, P., Ma, J., Wang, Y., Ma, L., and Huang, S. A context-aware method for top-k recommendation in smart tv. In *Asia-Pacific Web Conference* (2016), Springer, pp. 150–161.

[46] Mathison, S. Why triangulate? *Educational researcher 17*, 2 (1988), 13–17.

[47] Melville, P., Mooney, R. J., and Nagarajan, R. Content-boosted collaborative filtering for improved recommendations. *Aaai/iaai 23* (2002), 187–192.

[48] Meyer, K. How chunking helps content processing. *Nielsen Norman Group, Articles* (2016). Online at: https://www.nngroup.com/articles/chunking.

[49] Millen, D. R. Rapid ethnography: time deepening strategies for hci field research. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques* (2000), ACM, pp. 280–286.

[50] Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* (2018).

[51] Mittal, S., Agarwal, S., and Nigam, M. J. Real time multiple face recognition: A deep learning approach. In *Proceedings of the 2018 International Conference on Digital Medicine and Image Processing* (2018), ACM, pp. 70–76.

[52] Mori, M., MacDorman, K. F., and Kageki, N. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine 19*, 2 (2012), 98–100.

[53] Mortensen, M., Gurrin, C., and Johansen, D. Real-world mood-based music recommendation. In *Asia Information Retrieval Symposium* (2008), Springer, pp. 514–519.

[54] Morville, P., and Callender, J. *Search patterns: design for discovery.* " O'Reilly Media, Inc.", 2010.

[55] Mozer, T. F., Mozer, F. S., and Adams, E. B. System and method for controlling the operation of a device by voice commands, Aug. 10 2010. US Patent 7,774,204.

[56] Mróz, A. Filtration failure: On selection for societal sanity. *Kultura i Historia 34*, 2 (2018), 72–89.

[57] Newell, A., Simon, H. A., et al. *Human problem solving.* Prentice-Hall Englewood Cliffs, NJ, 1972.

[58] Oh, J., Sung, Y., Kim, J., Humayoun, M., Park, Y.-H., and Yu, H. Time-dependent user profiling for tv recommendation. In *2012 Second International Conference on Cloud and Green Computing* (2012), IEEE, pp. 783–787.

[59] Papadimitriou, A., Symeonidis, P., and Manolopoulos, Y. A generalized taxonomy of explanations styles for traditional and social recommender systems. *Data Mining and Knowledge Discovery 24*, 3 (2012), 555–583.

[60] Payne, J. W. Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational behavior and human performance 16*, 2 (1976), 366–387.

[61] Pazzani, M. J., and Billsus, D. Content-based recommendation systems. In *The adaptive web.* Springer, 2007, pp. 325–341.

[62] Preece, J., Rogers, Y., and Sharp, H. *Interaction Design: Beyond Human-Computer Interaction*, 4th ed. John Wiley & Sons, Inc, 2015. ISBN:111906600X, 9781119066002.

[63] Premack, D., and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences 1*, 4 (1978), 515–526.

[64] Pu, P., Chen, L., and Hu, R. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems* (2011), ACM, pp. 157–164.

[65] Pyae, A., and Joelsson, T. N. Investigating the usability and user experiences of voice user interface: a case of google home smart speaker. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (2018), ACM, pp. 127–131.

[66] Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence* (2009), AUAI Press, pp. 452–461.

[67] Rosenfeld, L., and Morville, P. *Information architecture for the world wide web.* " O'Reilly Media, Inc.", 2002.

[68] Russell-Rose, T., and Tate, T. *Designing the search experience: The information architecture of discovery.* Newnes, 2012.

[69] Samek, W., Wiegand, T., and Müller, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017).

[70] Sandvig, J. J., Mobasher, B., and Burke, R. Robustness of collaborative recommendation based on association rule mining. In *Proceedings of the 2007 ACM conference on Recommender systems* (2007), ACM, pp. 105–112.

[71] Sarwar, B., Karypis, G., Konstan, J., Riedl, J., et al. Analysis of recommendation algorithms for e-commerce. In *EC* (2000), pp. 158–167.

[72] Schwartz, B. *The paradox of choice: Why more is less.* Ecco New York, 2004.

[73] Shi, Y., Larson, M., and Hanjalic, A. Mining mood-specific movie similarity with matrix factorization for context-aware recommendation. In *Proceedings of the workshop on context-aware movie recommendation* (2010), ACM, pp. 34–40.

[74] Soares, M., and Viana, P. Tuning metadata for better movie content-based recommendation systems. *Multimedia Tools and Applications 74*, 17 (2015), 7015–7036.

[75] Soares, M., and Viana, P. The semantics of movie metadata: enhancing user profiling for hybrid recommendation. In *World Conference on Information Systems and Technologies* (2017), Springer, pp. 328–338.

[76] Soleymani, M., Koelstra, S., Patras, I., and Pun, T. Continuous emotion detection in response to music videos. In *Face and Gesture 2011* (2011), IEEE, pp. 803–808.

[77] Su, X., and Khoshgoftaar, T. M. A survey of collaborative filtering techniques. *Advances in artificial intelligence 2009* (2009).

[78] Syam, S. S., and Bhatnagar, A. A decision support model for determining the level of product variety with marketing and supply chain considerations. *Journal of Retailing and Consumer Services 25* (2015), 12–21.

[79] Takács, G., Pilászy, I., and Tikk, D. Applications of the conjugate gradient method for implicit feedback collaborative filtering. In *Proceedings of the fifth ACM conference on Recommender systems* (2011), ACM, pp. 297–300.

[80] Thomas, F., Baum, R. A., Hanes, D. H., and Main, J. M. Virtual hand based on combined data, Aug. 14 2018. US Patent App. 10/048,779.

[81] Todorovic, D. Gestalt principles. *Scholarpedia 3*, 12 (2008), 5345.

[82] Valenza, G., Nardelli, M., Lanata, A., Gentili, C., Bertschy, G., Paradiso, R., and Scilingo, E. P. Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis. *IEEE Journal of Biomedical and Health Informatics 18*, 5 (2014), 1625–1635.

[83] Van Meteren, R., and Van Someren, M. Using content-based filtering for recommendation. In *Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop* (2000), pp. 47–56.

[84] Vig, J., Sen, S., and Riedl, J. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent user interfaces* (2009), ACM, pp. 47–56.

[85] Wang, L., Meng, X., Zhang, Y., and Shi, Y. New approaches to mood-based hybrid collaborative filtering. In *Proceedings of the workshop on context-aware movie recommendation* (2010), ACM, pp. 28–33.

[86] Winoto, P., and Tang, T. Y. The role of user mood in movie recommendations. *Expert Systems with Applications 37*, 8 (2010), 6086–6092.

[87] Wioleta, S. Using physiological signals for emotion recognition. In *2013 6th International Conference on Human System Interactions (HSI)* (2013), IEEE, pp. 556–561.

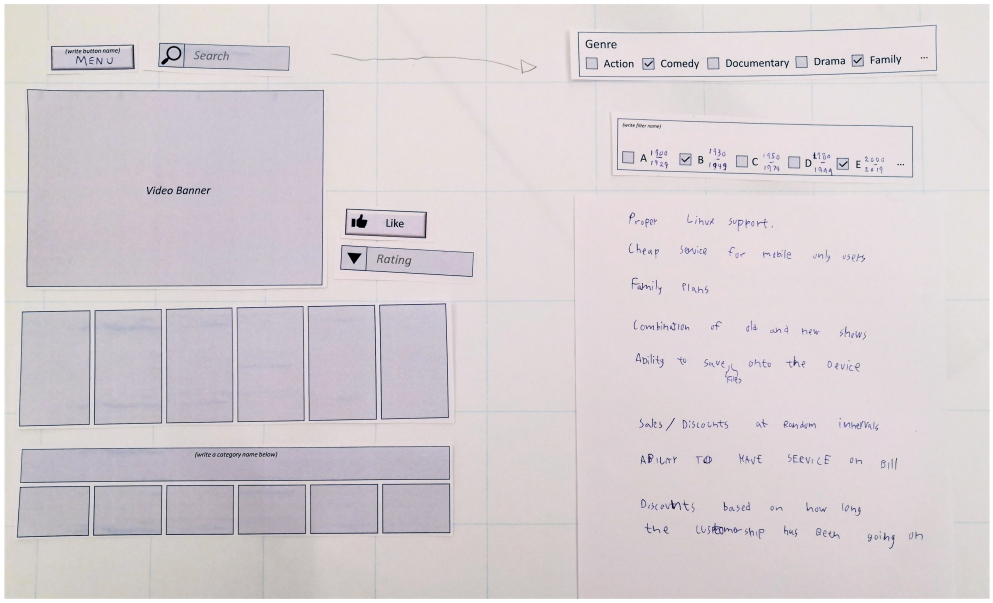
[88] Yu, H., Zheng, D., Zhao, B. Y., and Zheng, W. Understanding user behavior in large-scale video-on-demand systems. In *ACM SIGOPS Operating Systems Review* (2006), vol. 40, ACM, pp. 333–344.

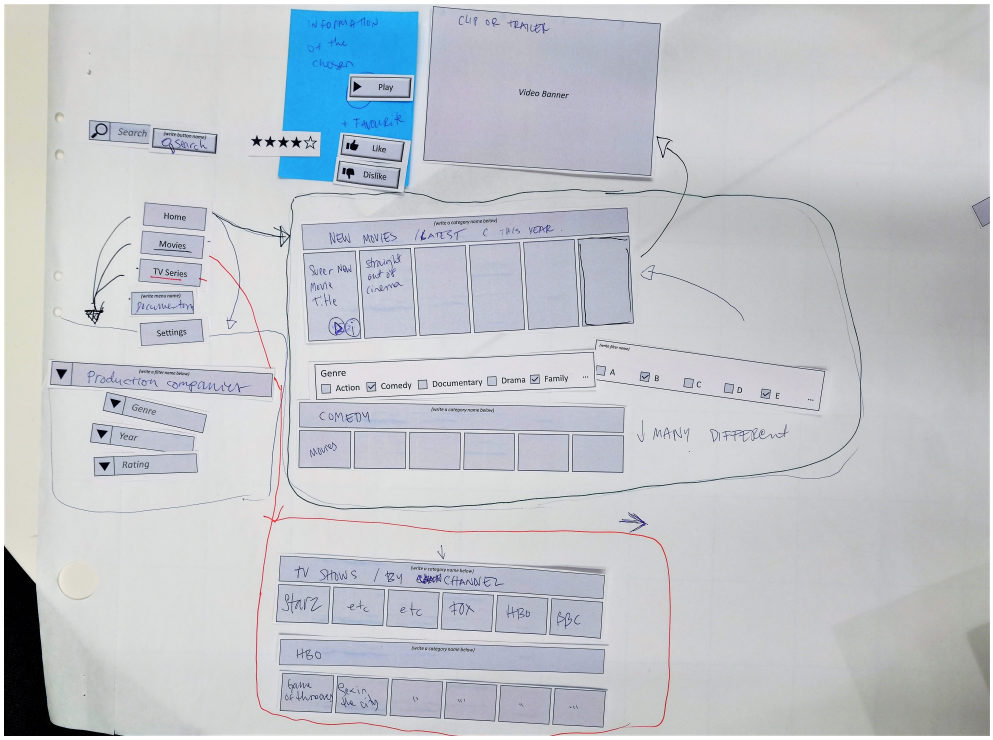
[89] Yu, Z., and Zhang, C. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (2015), ACM, pp. 435–442.

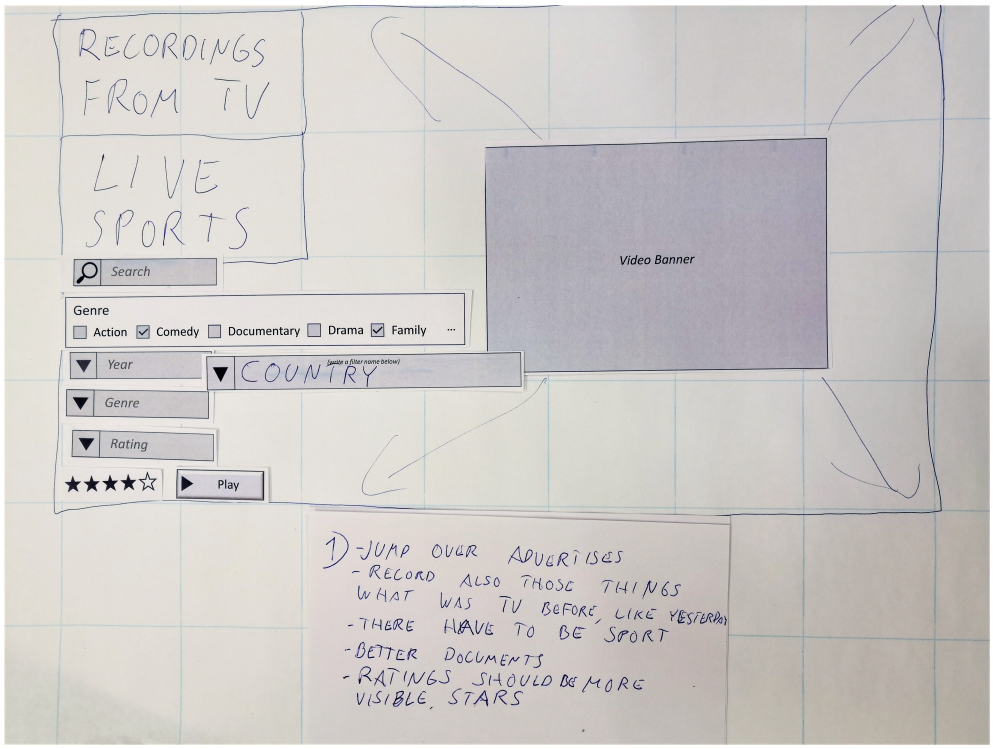
[90] Ziegler, C.-N., McNee, S. M., Konstan, J. A., and Lausen, G. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web* (2005), ACM, pp. 22–32.

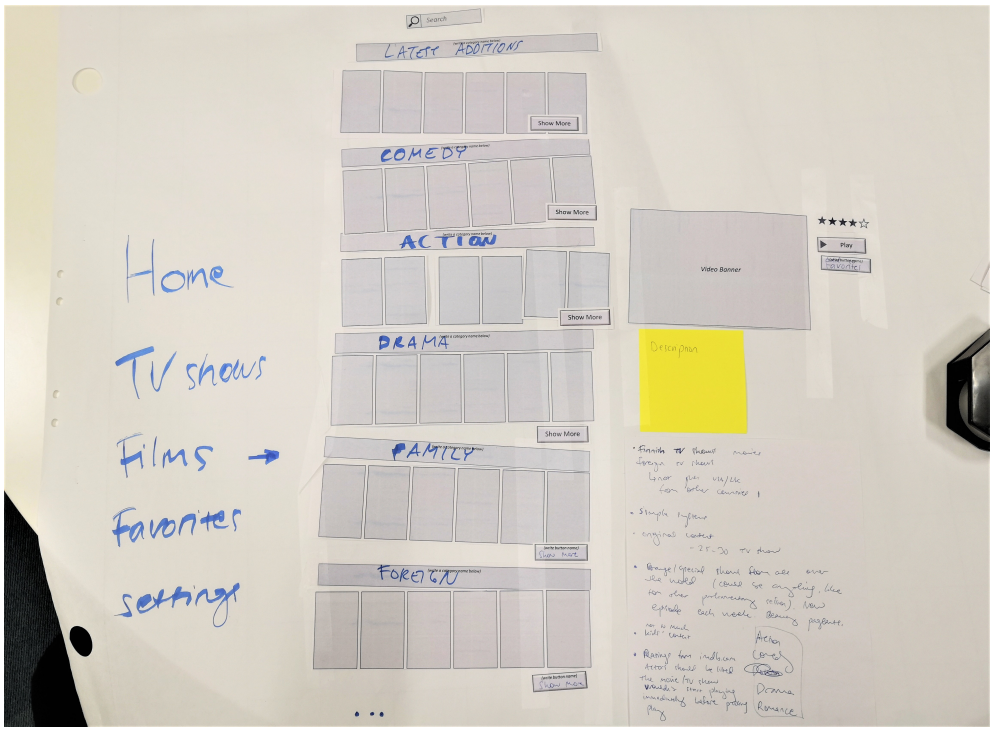
Appendix A

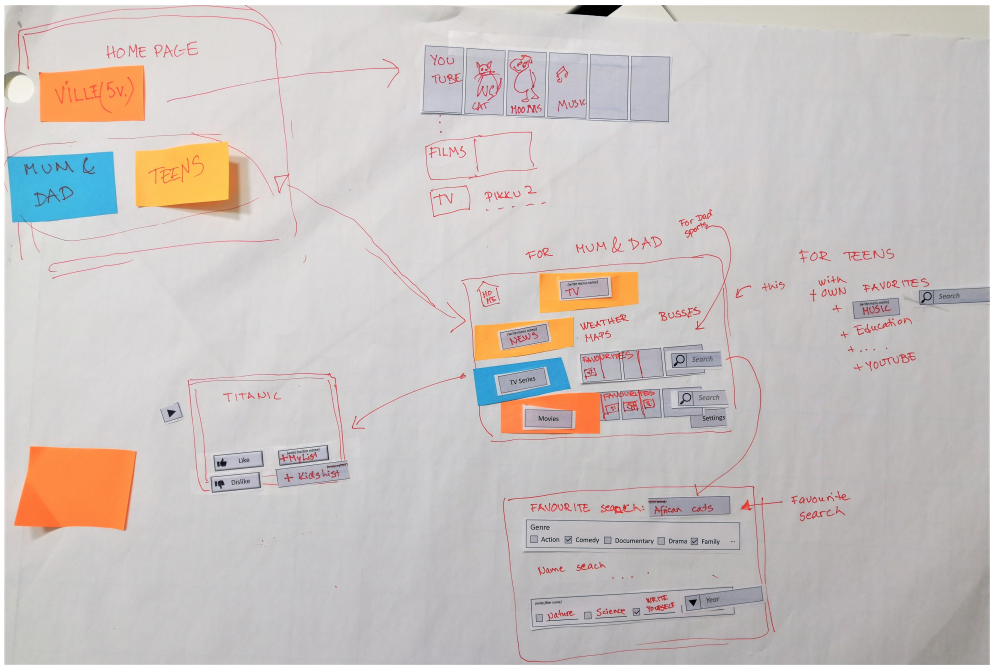
Appendix: Design Assignment

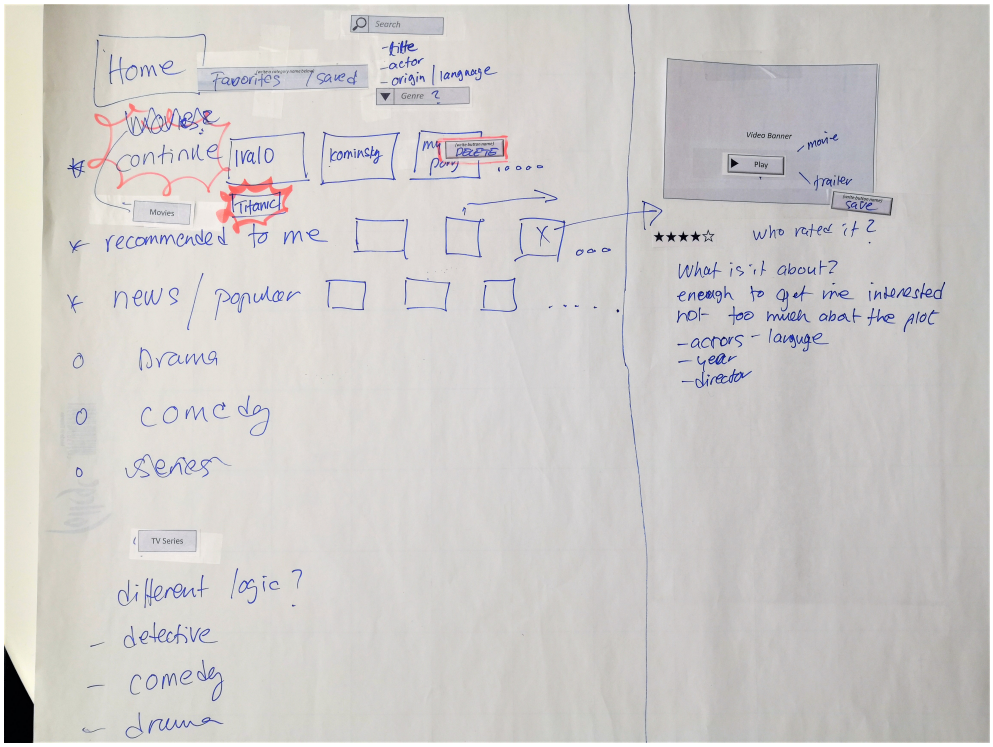


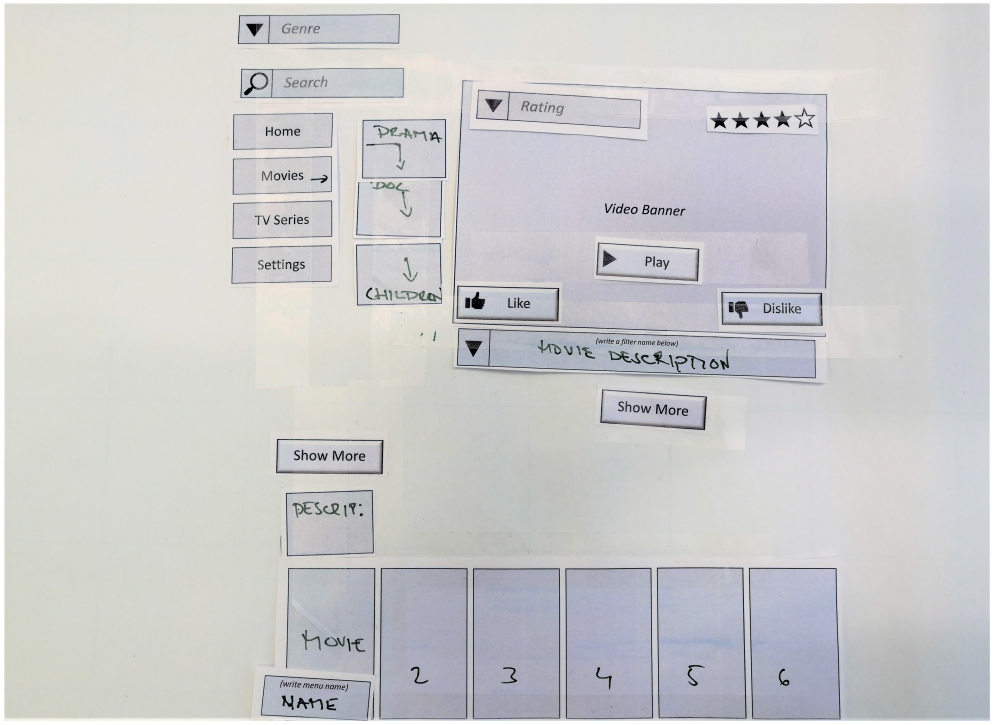

Play
Like
Dislike
Video Banner
Search
Home
Movies
TV Series
Settings
Production companies
Genre
Year
Rating
NEW MOVIES
Genre
Action
Comedy
Documentary
Drama
Family
COMEDY
TV SHOWS
HBO

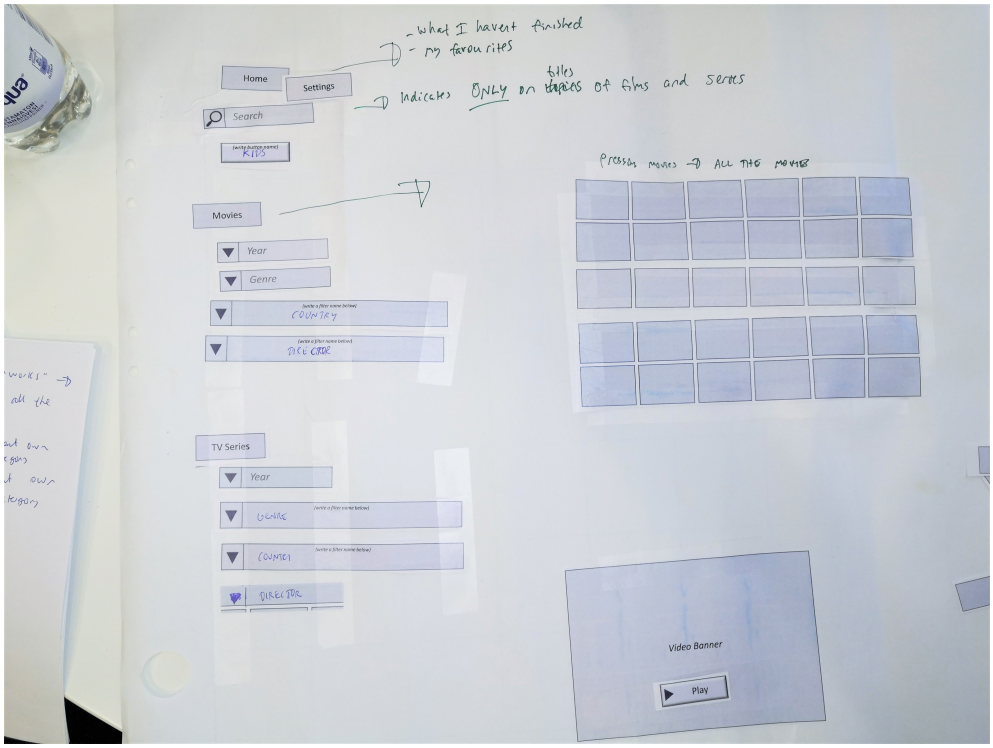

RECORDINGS FROM TV
LIVE SPORTS
Search
Genre
Action
Comedy
Documentary
Drama
Family
Year
COUNTRY
Genre
Rating
Play
Video Banner
1) -JUMP OVER ADVERTISES
- RECORD ALSO THOSE THINGS WHAT WAS TV BEFORE, LIKE YESTERDAY
- THERE HAVE TO BE SPORT
- BETTER DOCUMENTS
- RATINGS SHOULD BE MORE VISIBLE, STARS

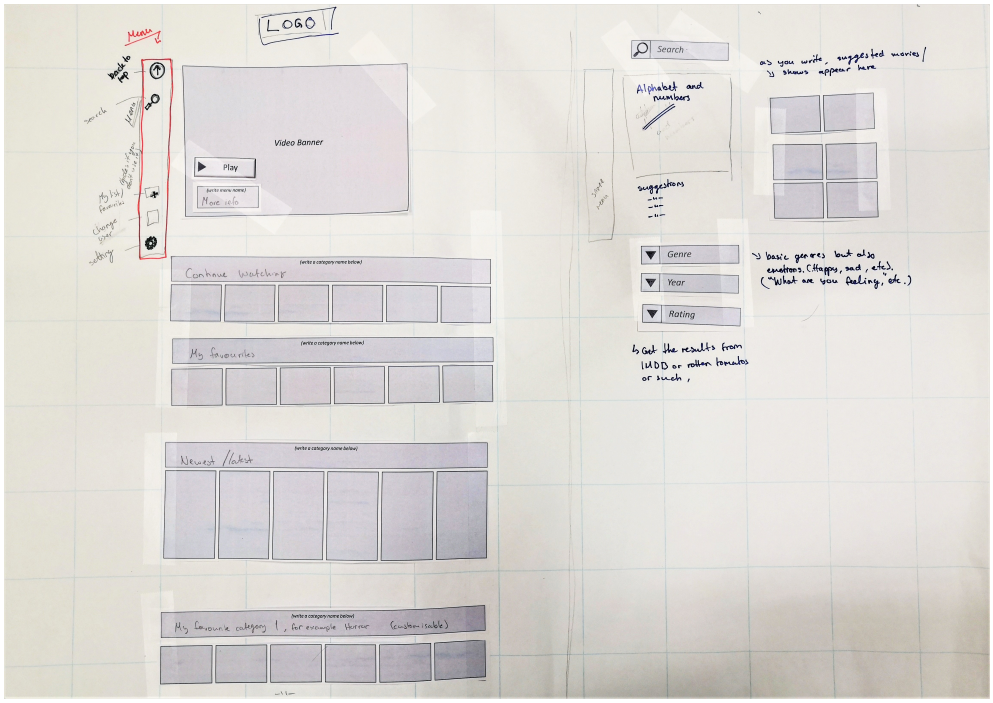

Search
LATEST ADDITIONS
Show More
COMEDY
Show More
ACTION
Show More
DRAMA
Show More
FAMILY
FOREIGN
Video Banner
Play
Description
Home
TV shows
Films
favorites
settings

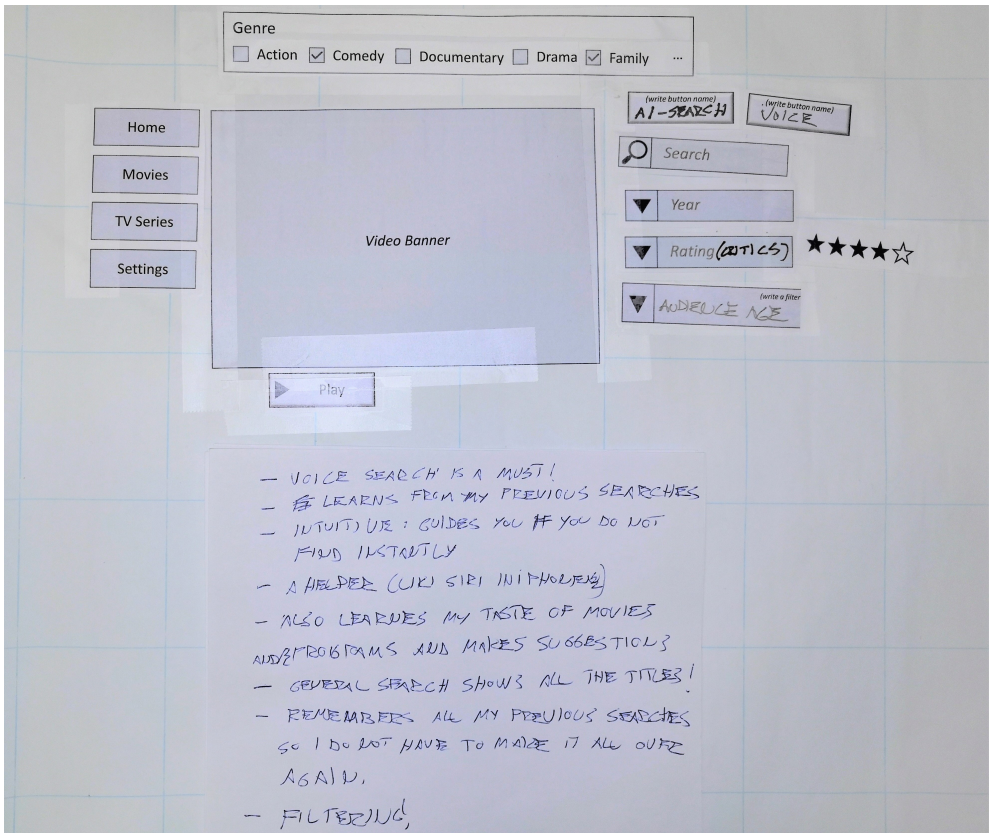

HOME PAGE
VILLE (5y.)
MUM & DAD
TEENS
YOU TUBE
MUSIC
FILMS
TV
PIKKU 2
FOR MUM & DAD
For Dad sports
this
TV
NEWS
WEATHER MAPS
BUSSES
TV Series
Movies
Search
Settings
FOR TEENS
with own
FAVORITES
MUSIC
+ Education
+ ...
+ YOUTUBE
TITANIC
Like
Dislike
+ kidslish
FAVOURITE SEARCH: African cats
Favourite search
Genre
Action
Comedy
Documentary
Drama
Family
Name seach
Nature
Science
Year


Search
Home
Favorites / saved
- title
- actor
- origin / language
Genre ?
Movies
continue
Titanic
recommended to me
news / popular
Drama
comedy
series
TV Series
different logic?
- detective
- comedy
- drama
Video Banner
- movie
Play
- trailer
save
who rated it?
What is it about?
enough to get me interested
not too much about the plot
- actors - language
- year
- director


Genre
Search
Home
Movies
TV Series
Settings
DRAMA
DOC
CHILDREN
Rating
Video Banner
Play
Like
Dislike
MOVIE DESCRIPTION
Show More
Show More
DESCRIP:
MOVIE
NAME
2
3
4
5
6


- what I havent finished
- my favourites
Home
Settings
Search
Indicates ONLY on titles pages of films and series
Movies
Year
Genre
COUNTRY
DIRECTOR
TV Series
Year
GENRE
COUNTRY
DIRECTOR
Video Banner
Play


LOGO
Video Banner
Play
More info
Continue watching
My favourites
Newest / latest
Search
Alphabet and numbers
suggestions
Genre
Year
Rating


Genre
Action
Comedy
Documentary
Drama
Family
...
Home
Movies
TV Series
Settings
Video Banner
Play
(write button name)
AI-SEARCH
(write button name)
VOICE
Search
Year
Rating (CRITICS)
(write a filter)
AUDIENCE
— VOICE SEARCH IS A MUST!
— IT LEARNS FROM MY PREVIOUS SEARCHES
— INTUITIVE: GUIDES YOU IF YOU DO NOT FIND INSTANTLY
— A HELPER (LIKE SIRI IN IPHONES)
— ALSO LEARNES MY TASTE OF MOVIES AND PROGRAMS AND MAKES SUGGESTIONS
— GENERAL SEARCH SHOWS ALL THE TITLES!
— REMEMBERS ALL MY PREVIOUS SEARCHES SO I DO NOT HAVE TO MAKE IT ALL OVER AGAIN.
— FILTERING!